# COMPUTATIONAL CHARACTERIZATION OF LONG NON-CODING RNAS

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR rerum naturalium
(Dr. rer. nat.)

im Fachgebiet

INFORMATIK

Vorgelegt

von M.Sc. Computer Science Rituparno Sen

geboren am 17. Mai 1988 in Coochbehar, Indien

Die Annahme der Dissertation wurde empfohlen von:

1. Prof. Dr. Peter F. Stadler, Universität Leipzig

2. Prof. Dr. Erich Bornberg-Bauer, Universität Münster

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 12.03.2021 mit dem Gesamtprädikat *cum laude*

# Bibliographic Description

| | |
|---|---|
| Title : | Computational Characterization of Long Non-coding RNAs |
| Type : | Dissertation |
| Author : | Rituparno Sen |
| Year : | 2020 |
| Discipline : | Computer Science |
| Language : | English |
| Keywords : | lncRNAs, NGS, annotation, lncRNA function, machine learning |

The following publications form the basis of the thesis:

**Rituparno Sen**, Jörg Fallmann, Maria Emília M. T. Walter, and Peter F. Stadler. *Are spliced ncRNA host genes distinct classes of lncRNAs?* Theory in Biosciences, 2020. doi: 10.1007/s12064-020-00330-6

**Rituparno Sen**, Gero Doose, and Peter F. Stadler. *Rare splice variants in long non-coding RNAs.* Non-coding RNA, 3(3):23, 2017. doi: 10.3390/ncrna3030023

*42 is the answer to the ultimate question of life, the universe and everything.*

Deep Thought

# Abstract

In a cell, the DNA undergoes transcription to form mature transcripts, some of which in turn undergo translation to form proteins. Although over 85% of the human genome is transcribed, it comprises only about 2% protein-coding genes, the rest being non-coding. One of the non-coding gene elements, called long non-coding RNAs (lncRNAs), are emerging as key players in various regulatory roles in the human genome. The generally accepted theory posits lncRNAs to be over 200 nucleotides long and to be able to grow over 10 kilobases, bearing a similarity with mRNAs. The majority of lncRNAs undergo alternative splicing and are weakly polyadenylated in combination with complex secondary structures. Among the annotated lncRNAs, so far it has been only a meagre portion for which functional roles have been detected, while functions of the vast majority remain to be discovered. Observed functional roles include thus far gene expression regulation through various mechanisms at transcriptional and post-transcriptional levels. With the advent of next-generation sequencing (NGS) and advances in RNA sequencing technology (RNA-Seq), it is easier to reconstruct the transcriptome by extracting information about the splicing machinery. RNA-Seq has helped consortia like GENCODE, ENCODE, and others to curate their annotation catalogues. In this PhD thesis, certain aspects of the human lncRNA transcriptome will be explored, such as the challenges in lncRNA annotation. Those challenges stem from the lack of signals that are common in mRNAs and make them easier to detect, for instance signals of ORFs and transcription start sites. Concurrently, owing to a lack of understanding of the connection between sequence and function, lncRNAs have been typically annotated based upon their location in relation to mRNAs and their functions have been predicted through a guilt-by-association approach.

In the first part of the PhD research work, the splice junctions in the lncRNA transcriptome were mapped in an attempt to explore the isoform diversity of lncRNAs by using sequencing data from B-cell lymphoma. In this phase of the research work, multiple junction-spanning reads from the sequencing data with a very large read depth were found to represent the splice junctions. Using GENCODE v19 as a reference it was found that the human transcriptome harbours a large number of rare exons and introns that have remained unannotated. Concomitantly, it can be inferred that the current human transcriptome annotation is confined to a very well-defined set of splice variants. However, although the isoforms are well-defined, the same cannot be said about their biological functions and it remains to be explored why the processing machinery of lncRNAs is restricted to a set of very few splice sites.

In the human genome, small regulatory RNAs like miRNAs and small nucleolar RNAs (snoRNAs) overlap with lncRNAs in their genomic loci. To further understand

the human transcriptome, in the second part of the PhD research work, a study was undertaken in an attempt to distinguish the miRNA and snoRNA hosting lncRNAs from the lncRNAs that did not have any overlaps with the smaller RNAs. To this end, machine learning techniques were implemented on curated datasets employing features inspired by a few of the prevalent features used in published lncRNA detection tools encompassing not just sequence information, but also secondary structure and conservation information. Classification was attempted through supervised as well as unsupervised learning approaches; random forests for the former, PCA and $k$-means for the latter. In the end, the three RNA classes could not be separated with certitude, especially when the hosted RNA was not supplied to the classifier, however, this lack of detectable association can be confirmed to be of biological interest. It suggests that the function of host genes is not closely tied to the function of the hosted genes at least in this case. Nevertheless, understanding the dynamics of snoRNA and miRNA host genes can improve the knowledge of functional evolution of lncRNAs, as the fact that the smaller RNA genes are conserved makes it comparably easier to trace the host lncRNAs over much larger evolutionary timescales than most other lncRNAs. With the accelerated availability of sequencing techniques it can be expected that expanded investigation into conservation patterns and host gene functions will be possible in the near future.

**keywords**: lncRNAs, NGS, annotation, lncRNA function, machine learning

# Acknowledgements

First and foremost, I would like to thank Prof. Peter Stadler for the opportunity to work with him and learn from him. It started with an email all those years ago, leading to the reception of a scholarship and my subsequent travel to Germany. His guidance has been decisive during the period of the doctoral study and at its culmination, I can only look back at the experience I've gained, which will be invaluable for the future. Thank you, Peter.

Some extraordinary support that I've received in the past two years comes in the form of a jolly gentleman, Dr. Jörg Fallmann. He has been encouraging me, guiding me, disentangling the future path of work - the list goes on. I wonder whether he slept enough in the last few days before the submission of my thesis with all my at times odd questions.

I would also like to take the opportunity to thank the DAAD for the financial and official support that enabled me to a smooth beginning in Germany.

The next cheers go out to the warriors of the group: Petra Pregel, for getting me through the bureaucratic quagmire successfully, and Jens Steuck, coaxing the workstation to get back up and running when it mysteriously failed to do so.

Dr. Deisy Gysi richly deserves a special note of thanks for accompanying me through most of my journey and especially right at the end, with words of encouragement and of technology.

I can't possibly have carried on without the presence of the ever-present friends Wilmer and Angie and the MIA Eugenio. Take a bow, ye faithless lot.

To all the people that have made my stay in Leipzig so much more exciting and enriching - Cheers!

For being a pillar through the hard times, for cheering me up, for accepting my unreasonable demands - I would have spiralled into insanity, if not for the presence of Tjorven Bienfait. May the forthcoming days be brighter for us.

The final note of thanks will go out to the biggest support system of my life - my parents. Although being 8,000 km away, they have always been there, despite my occasional inattentiveness. Thank you.

# Contents

# List of Figures

# Tables

# I

# Introduction

# 1

# Motivation

Long non-coding RNAs (lncRNAs) are emerging as key players in various regulatory roles in the human genome, which are similar to mRNAs length-wise, however, are not identical. The generally accepted theory posits lncRNAs to be over 200 nucleotides (nts) long and to be able to grow to over 10 kilobases. LncRNAs do not code for proteins, as they lack open reading frames (ORFs), although there is evidence that some are translated to form peptides. A majority of lncRNAs undergo alternative splicing and are weakly polyadenylated. Similar to mRNAs, lncRNAs are primarily transcribed by RNA polymerase-II (pol-II) in eukaryotes, however, some are transcribed by RNA pol-III. In contrast, some plant lncRNAs are transcribed by RNA pol-IV and pol-V. Most lncRNA genes are located in the vicinity of promoter regions of coding genes. Additionally, some lncRNA transcripts have a $m^7G$ cap at their 5*'* end and they also tend to possess complex secondary structures. Among the annotated lncRNAs, so far it has been only a meagre portion for which functional roles have been detected, while functions of the vast majority remain to be discovered. Observed functional roles include thus far gene expression regulation through various mechanisms at transcriptional and post-transcriptional levels that make them an excellent case study[1–3].

With the advent of RNA sequencing technology (RNA-Seq), it has become easier to reconstruct the transcriptome by extracting information about the splicing machinery, thereafter detecting exons and eventually transcripts. RNA-Seq has helped consortia like GENCODE, ENCODE, and others to curate their annotation catalogues. In this PhD thesis, certain aspects of the human lncRNA transcriptome will be explored, such as the challenges in lncRNA annotation. Those challenges stem from the lack of signals that are common in mRNAs and make them easier to detect, for instance, signals of ORFs and transcription start sites. Concurrently, owing to a lack of understanding of the connection between sequence and function, lncRNAs have been typically annotated based upon their location in relation to mRNAs and their functions have been predicted through a guilt-by-association approach. In contrast to the more abundant mRNAs,

lncRNAs are less abundant and less stable, and most of the genes are lowly-expressed, which in turn leads to difficulties in their identification[4;5].

In the first part of the PhD research work, the splice junctions in the lncRNA transcriptome were mapped in an attempt to explore the isoform diversity of lncRNAs by using sequencing data from B-cell lymphoma[6]. Typically, lncRNAs possess multiple exons and an affinity towards two-exon transcripts, and undergo splicing to form different transcripts. It is reported that the lncRNAs have one dominant isoform and the alternative isoforms do not have similar expression levels, whereas mRNAs produce at least two dominant isoforms. To explore the mechanisms of the splicing machinery that lncRNAs possess, transcriptome data with a large read depth was employed in this phase of the research work, and concurrently, multiple junction-spanning reads from the sequencing data were found to represent the splice junctions. GENCODE v19 was used as reference, as it included the maximum number of overlaps with the sequencing data. In the process, it was found that the human transcriptome harbours a large number of rare exons and introns that have remained unannotated. Since the read depth of the lymphoma data was very large (around $10^{10}$ reads), those splice variants could be detected. A near perfect saturation of the average number of splice junctions per gene was also reached. Concomitantly, it can be inferred that the current human transcriptome annotation is confined to a very well-defined set of splice variants. However, although the isoforms are well-defined, the same cannot be said about their biological functions and it remains to be explored why the processing machinery of lncRNAs is restricted to a set of very few splice sites.

In the human genome, small regulatory RNAs like miRNAs and small nucleolar RNAs (snoRNAs)[7] overlap with lncRNAs in their genomic loci. To further understand the human transcriptome, in the second part of the PhD research work, a study was designed in an attempt to distinguish the miRNA and snoRNA hosting lncRNAs from the lncRNAs that did not have any overlaps with the smaller RNAs. To this end, several machine learning techniques were implemented. In recent times, numerous tools have been developed that can successfully distinguish mRNAs from lncRNAs, detect miRNAs, or even separate miRNAs from snoRNAs. The features used for the current classification problem were inspired by a few of the prevalent features used in those tools encompassing not just sequence information, but also secondary structure and conservation information. The features used were $k$-mer profiles, Fickett score, conservation scores, pairing probability between nucleotides, and GC content. A random forest classifier was trained on the curated datasets as part of the supervised learning approach, whereas PCA and $k$-means clustering were employed as part of the unsupervised training approach. A convolutional neural network (CNN) was also constructed, but did not return any credible results, as the scope of the data was very limited to train a deep learning algorithm. In the end, the three RNA classes could not be separated with certitude, especially when the hosted RNA was not supplied to the classifier, however, this lack of detectable association can be confirmed to be of biological interest. It suggests that the function of host genes is not closely tied to the function of the hosted genes. In contrast, protein-coding host genes of snoRNAs contribute to the maturation of the ribosome. The lack of common, class-specific features for host genes together with their usually very poor sequence conservation suggests that they may even have acquired different functions in different lineages. Nevertheless, understanding the dynamics of snoRNA and miRNA host genes can improve the knowledge of functional

evolution of lncRNAs, as the fact that the smaller RNA genes are conserved makes it comparably easier to trace the host lncRNAs over much larger evolutionary timescales than most other lncRNAs. With the accelerated availability of sequencing techniques it can be expected that expanded investigation into conservation patterns and host gene functions will be possible in the near future.

# 2

# Outline

This PhD dissertation comprises three main parts.

The first part is further subdivided into two chapters. Chapter 3 is dedicated to an overview of long non-coding RNAs (lncRNA) and their regulatory mechanisms according to established literature. The recent developments in next-generation sequencing (NGS) along with challenges of lncRNA annotation are explored. Many lncRNAs have been implicated in various diseases, which will be detailed next. Following which, a brief introduction on machine learning can be found in Chapter 4. Machine learning techniques are being widely used in computational biology. They are also being implemented in detection, analysis, and classification of long non-coding RNAs. The tenet of machine learning is briefly discussed in this chapter besides a few well-established algorithms along with several evaluation metrics and challenges.

The second part is subdivided into four chapters and will focus on the results of the transcriptomic classification study, that has been carried out within the period of the doctoral programme. The first chapter (Chapter 5) explains a study of splice junctions in the human transcriptome using sequencing data from B-cell lymphoma and a reference annotation catalogue. Chapter 6 outlines a review on several machine learning techniques used in lncRNA detection, ranging from supervised approaches to deep learning. Chapter 7 describes a study exploring the relationship between the lncRNA genes and miRNAs and snoRNAs, whose genomic loci overlap. The aim of this study was to find out using machine learning if lncRNAs hosting smaller RNAs and lncRNAs that do not are distinct classes of non-coding RNAs. Furthermore, if a relationship exists that would also throw some light towards functional classification of lncRNAs. Chapter 8 explores the results of the classification study.

The third and the final part focuses on the conclusion of this doctoral study.

# II

# Biological and technical background

# 3

# Long non-coding RNAs and transcriptomics

In this chapter, a background on long non-coding RNAs (lncRNAs) and transcriptomics will be provided. LncRNAs, the first of which was discovered in the 1980s, although mis-classified as an mRNA, are in the continuous process of discovery. The lncRNAs have emerged as important players in gene regulation and have been found possessing vital roles in various biological processes and diseases. With the advent of high throughput sequencing technologies, the whole of human genome was mapped in the early 2000s. One of the RNA categories that benefited the most was lncRNAs: the availability of sequencing data of the human genome spurred on research into lncRNAs to gain more functional insights. The Human Genome Project[8] mapped around 3 billion base pairs (bp) encoded in 23 pairs of chromosomes (22 pairs of autosomes and one pair of sex chromosomes, X and Y) in the human genome and that inspired the project ENCODE[9;10], which mapped regions of transcription, transcription factor association, chromatin structure and histone modification in the human genome. It was realised that around 1.2-1.5% of the human genome codes for proteins, with the rest being populated by non-coding RNAs[10]. There are about 20,000 protein-coding genes currently annotated, whereas the number of lncRNAs is estimated to be varying depending on the annotation databases and pipelines[5;11]. A brief discussion about the next-generation sequencing (NGS) technologies will be explored first to understand the underlying challenges of lncRNA annotation, following which the biogenesis and regulatory mechanisms of lncRNAs will be described.

## 3.1 Transcriptomics

### 3.1.1 Next-Generation Sequencing

The Human Genome Project was the result of a plethora of technological innovations encompassing fields of chemistry, molecular biology, engineering, and software that led to the inception of fast, automated DNA sequencing machines with the technology being known as Next-Generation Sequencing (NGS)[8;12–14]. The method was first described by Sanger et al.[15], who mixed the four native deoxynucleotides with dideoxynucleotides to obtain fragments that were nucleotide-specific[14]. As a result of a further polymerase chain reaction (PCR), the occurring strand elongation was studied and the strands were separated on polyacrilamide gels. The separated fragments were then labelled as bases through laser excitation and spectral emission analysis[13;14]. The method was upgraded by radio-labelling dATP (a substrate of DNA polymerase used for synthesis) with fluorescent labelled primers, which made the whole process automated. The rate at which DNA strands could be sequenced received a further boost when slab gels were eliminated in favour of capillaries, where base separation was executed through electrokinetic injection and provided single nucleotide resolution[14].

All these methods were superseded by NGS. DNA sequences undergo **fragmentation**, *i. e.* they become a part of a library of fragments with trailing adapters (synthetic DNA to boost polymerase content in those fragments); the to-be-sequenced DNA fragment being known as a **template**. It is done through utilising DNA ligase and the amplified fragments result in a single focus that determines the sequencing data for a fragment[14]. In contrast to staggered nature of Sanger sequencing, massively parallel NGS operates continuously by performing the sequencing and detection of each nucleotide (or fragment), which enables this technology to handle billions of reaction foci. Enormous amounts of data are generated; however, the error rates can be high[13]. The scalability and throughput of NGS have enabled generation of sequences for whole genomes of many organisms, not least of human[9;13;14]. The underlying alignment techniques of NGS constitute *short* ***reads*** (75-300 bp long) and *long* ***reads*** ($>$ 1000 bp), that power sequence analysis leveraging a high-quality reference genome or *de novo* construction of genome; **read** being a sequence of DNA bases[13;14]. Short-read sequencing use either sequencing by ligation or sequencing by synthesis. Bead-based, solid-state, and DNA nanoball generation are different strategies to create clonal templates. Long-read sequencing utilise single-molecule sequencing and synthetic approaches to construct long reads from short reads[13]. Two types of sequencing strategies are in use: single-end and paired-end. Single-end sequencing entails sequencing of a DNA template only from one end. On the other hand, paired-end sequencing can be described as having an adapter sequence each on either end of a DNA fragment ($\leq$ 1 kb) and extracting two different reads based upon priming of both adapters. This ensures a more accurate placement of the read than from a single-end read of the same length. It is worth noting that the forward and reverse reads may or may not overlap[13]. Additionally, paired-end reads allow for insertion of long sequences between the reads. Concomitantly, DNA fragments longer than 1 kb are joined together at ends using one adapter, instead of two, in mate-pair sequencing, an extension of the paired-end approach. The fragment is then processed and two reads are obtained and can be aligned to a reference genome.

However, the distance between the two reads is longer in comparison with the paired-end strategy. In a combination of paired-end and mate-pair approaches, longer range can be achieved, since reads can be distant from each other, and paired-end reads can align complicated regions with mate-pair reads providing the scaffold[14].

However, read length supported by an instrument leads to noise in the sequencing process and sequencing errors. The signal-to-noise ratio is determined by sequencing a reference set of genes and aligning it to a high-quality reference genome and it is referred to as the **error model** of an instrument. Coverage biases, insertion and deletion errors as well as errors caused by enzymatic amplification during library preparation contribute to the error model[14].

454 pyrosequencing was the first NGS instrument developed which used bead-based sequencing deploying a charge-coupled device camera to capture bioluminescence signals[13]. The first system developed to use reversible dye terminators in enzymatic sequencing of fragments was Solexa 1G which could process 25 bp single-end reads and was acquired and upgraded by Illumina to 150 bp paired-end reads using the HiSeq 2000[16]. Ion Torrent introduced sequencing by detection of pH change as hydrogen ions are released during nucleotide incorporation through the Ion S5 platform[17]. Single-molecule sequencing is a method of focusing on detection of a single molecule by injection of fluorescent labelled nucleotides combining nanotechnology with molecular biology. The zero-mode waveguide (ZMW) was the technology developed and deployed by Pacific Biosciences[14;18]. Sequel, developed by the company, can obtain sequences of average 10 kb read length and have high-throughput like their other instruments, although with a high single-pass sequence error, which can be decreased[12;19]. A single-molecule sequencer was also developed by Oxford Nanopore, called MinION, which was remarkable for its size, albeit with high error rates. The MinION sequences a double-stranded DNA either in one direction along either strand (1D read) or in both directions generating a consensus sequence (2D read)[13;20]. PromethION and GridION are two other parallel sequencing platforms developed by the company utilising multiple flow cell (disposable contents of sequencing) stacking[12;21;22]. Illumina developed several platforms targeting different requirements: MiSeq, NextSeq, MiniSeq, NovaSeq, and iSeq. NovaSeq can provide greater sequencing depth with moderate read lengths, whereas PacBio and Oxford Nanopore platforms can provide longer read supports as part of third-generation sequencing (TGS)[12]. The Illumina CRT system has been used for whole genome sequencing[13], whereas the 10X Genomics emulsion-based system and the Illumina synthetic long-read sequencing platform are the two systems available for generating synthetic long-reads[13].

### 3.1.2 RNA Sequencing and transcript reconstruction

The impact of the applications of RNA Sequencing technology (RNA-Seq) has been felt in several areas of molecular biology: from the splicing machinery of mRNAs to the regulation of gene expression by non-coding RNAs[20;23;24]. Most of the research has been carried out with Illumina short-read sequencing platforms resulting in comparatively faster and more accurate results than obtained with older microarray-based approaches. Concurrently, differential gene expression (DGE) assays have been preferably created

with an Illumina short-read sequencing platform[20]. 150-200 bp cDNA fragments are generated to construct the library before being analysed, which contains 20-30 million reads per sample. However, although short-read sequencing is cheaper and easier to implement than microarrays, that may fall short for whole-transcriptome analysis[20;25]. Multi-mapped reads - reads from homologous regions that cannot be mapped to the transcriptome unambiguously - and presence of isoforms of genes also need to be considered[20]. Long-read sequencing can eliminate the shortcomings of its counterpart by reading full length cDNA transcripts converted from mRNAs to identify isoform diversity. Additionally, splice junction detection is comparatively more reliable[26]. Oxford Nanopore provides a platform where mRNA transcripts can be directly sequenced without the need of conversion to cDNA and the process is known as dRNA-Seq[27;28]. Long-read sequencing is hampered by low throughput in comparison with short-read sequencing. The Illumina platform for the latter can generate $10^9 - 10^{10}$ short reads, whereas PacBio can attain $10^6 - 10^7$ long reads[20]. Despite accruing higher costs, long-read sequencing can detect isoforms previously undetected by short-read sequencing approaches, especially because of the read lengths. However, the error rates of long-read methods are high and generated data contain insertion and deletion errors, although some errors can be corrected by increasing the read depth[29;30].

RNA-Seq provides a technique to analyse polyadenylation, alternative splicing and alternative promoter usage[20]. In gene-level expression detection experiments, a tag read is generated for every fragment and counted that come from 3 ' ends of mRNAs and are unique to a transcript[31]. The 5 ' ends are analysed for transcription start site mapping and mostly executed by cap analysis for gene expression (CAGE)[32]. RNA-Seq experiments generally have high duplicate rates, which are thought to be artefacts. Single-end and paired-end sequencing approaches offer either the 3 ' or 5 ' end to be selected for the former, or both for the latter. Paired-end sequencing is preferred to have wider nucleotide coverage[20].

To analyse the data that has been generated by the sequencing experiments, the reads (in FASTA/Q format) are mapped to a reference transcriptome or genome to generate the genomic coordinates. This step allows for interpretability of splice junctions (detection of exons and introns), since the cDNA reads span exon boundaries[20]. A splice junction is described to be canonical when there are dinucleotides GT and AG detected at the donor and acceptor sites, respectively. The splice junctions can be accepted as canonical for GC/AG and AT/AC pairs, too; however, if any other pairs are detected, the splice site is called non-canonical[33]. Entire genome assembly is economically expensive, but since transcription constitutes only a partial fragment of the genome, RNA transcript reconstruction is relatively cheaper[34]. There are two tried and tested approaches: i) the reads are aligned to a reference genome and the transcript model is deduced, and ii) in the absence of a reference genome, contiguous transcript sequences are assembled in *de novo* reconstruction strategy to deduce possible splice junctions[34–37] In order to perform the alignment, tools such as TopHat2[38] and segemehl[39] are used. The tools can also assemble sequenced reads into transcripts in case of *de novo* reconstruction, however, they are not adequate, since they are plagued with alignment issues when, for example, there are sequencing errors or repeats and similar regions across copies of genes[40]. To perform transcript reconstruction, features such as exon and transcript identification, coding content, and understanding and estimating expression levels of the genes predicted are required[34]. Typically, the RNA-Seq reads (around 200 million,

75-150 bp) are normally assembled in a single set, which are then aligned to the reference genome (when there is one available) or realigned to the assembled transcripts in *de novo* reconstruction[34;41]. The exonic stretches of the transcripts are detected by identifying individual exons in the reads and is normally carried out by identifying the transcription start and end sites, however, there can be inaccuracies in determining them during sequencing resulting in false detection of exons[35;42;43]. Concordantly, transcripts can also be detected using the same machinery[35]. Concurrence in positioning of 5' and 3' ends between predicted and annotated transcripts can sometimes be hard to achieve, which can lead to misidentification of splice sites. Additionally, accurate detection of translation start and end sites give way to proper identification of coding exons, given enough read depth[35]. Similarly, introns are detected based upon the underlying read alignments and can be correctly identified through observing the overlaps with known splice sites[26;35]. For example, some reads can span two or more exons, *i. e.* multiple splice sites, because the mean exon length for many eukaryotic organisms is less than 200 bp[41]. Subsequently, expression levels of transcripts are estimated and are normalized by read depth and length (RPKM or FPKM), which can be affected by incomplete transcript models, where known and estimated expression levels differ; divergence in concurrence of exon distribution being a cause behind this[35;44].

Following mapping the reads onto a reference transcriptome to extract coordinates, they are assigned to transcripts or genes. To this end, the number of overlaps of reads with known transcripts are recorded based upon the abundance of the reads. This is useful for detecting isoforms, too. With a short-read approach, an estimation is done because not all reads will span splice junctions and cannot be assigned to an isoform unambiguously[45]. CuffLinks is one of frequently used quantification tools[46]. An expression matrix is created following quantification, containing transcript or gene names and read counts (or estimates)[20]. The differences in read depth, expression patterns, and technical biases can be accounted for by filtering the read counts of the quantified transcripts or genes, which can also lead to more accurate detection of differential expression of genes[47]. Expression matrices are also normalized based upon assumptions that most gene expression levels remain the same across replicate groups and mRNA levels of the sample groups are similar[20]. Concurrently, differential expression modelling of the transcripts or genes are carried out to detect the features that may have changed the expression. Tools like CuffDiff[48] or DESeq2[20] are used for this purpose. Finally, transcript reconstruction also depends upon the completeness of the reference genome annotation of the target organism[34;35;41].

Apart from sequencing regions from whole tissues, single-cell sequencing is also being considered for experiments to detect the full complement of cell types in an organism or tissue[20]. Endeavours like The Human Cell Atlas and NIH Brain Initiative attempt to harness the full extent of this technology[49;50]. To understand spatial information concerning the relationship between gene expression and cellular context, spatially restricted cells are isolated by laser-capture microdissection (LCM) or through barcoding RNAs. RNA-Seq can also be deployed to capture RNA dynamics by mapping TSSes and quantifying newly transcribed RNA (nascent RNA)[20]. So far, transcription divergence at promoter regions and gene regulation through active RNA Pol-II being paused in the proximity of promoters have been discovered, which show the regulatory roles nascent RNAs play[20;51]. Active translation can also be measured using polysomal profiling and ribosome footprinting by detecting the ribosomes present in a mature

mRNA transcript[52]. Besides that, RNA-RNA interactions and RNA-protein interactions can also be detected using RNA-Seq[53]. ChIP-Seq (chromatin immunoprecipitation and sequencing) is the usual method to detect and analyse binding sites of DNA-associated proteins[54]. For RNAs, RIP-Seq (RNA immunoprecipitation and sequencing) have been used, followed by a UV crosslinking strategy photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP), which attains nucleotide resolution and stabilises RNA-protein binding, without crosslinking protein-protein interactions[20;55].

### 3.1.3 lncRNA annotation

NGS has enabled researchers to identify lncRNA genes in almost all eukaryotes[4]. To leverage the information gained from lncRNA transcriptome, annotation is also fundamentally necessary to understand functional roles of lncRNAs. Annotation indicates cataloguing gene loci with genomic coordinates of transcripts and exons; they possess hierarchical information of overlapping transcripts and shared exons. Two different strategies are in place for annotation: automated and manual. Transcriptome assemblies are used for automated annotation rendering it to be fast, but often inaccurate, whereas manual annotation is performed by hand following RNA evidence and protocols and is slower, but has higher accuracy rates[4]. LncRNA genes have typically been annotated as genic or intergenic, depending upon whether the loci intersect with protein-coding genes, since there is a lack of understanding of the connection between sequence and functions of lncRNAs. It is challenging to annotate lncRNAs and it can be executed only through physical transcriptomic information. They have low expression levels, which are bound to be not substantiated or even missed in transcriptomic data, together with expressed sequence tag (EST) and CAGE data[4;56;57]. Unlike protein-coding genes, lncRNAs do not possess any concrete signals in forms of ORFs or TSSes, which deter smooth identification. Furthermore, identification through sequence similarity in orthologues or paralogues is difficult, since lncRNAs are weakly conserved[58]. Lowly-expressed transcripts pose a challenge to measure abundance of transcripts from RNA-Seq reads[59]. Accurate identification and handling of poly(A) tails is required as they affect estimated expression level[4]. For CRISPR-Cas screens, it is important to identify lncRNA TSSes for accurate targetting of promoter regions[60]. For genome-wide association studies (GWAS), which rely on information about trait-associated mutations, and identification of lncRNA biomarkers, accurate transcript reconstruction is imperative[4].

Since non-coding RNAs are expressed at lower levels than protein-coding genes, lncRNA exons have also been detected to be less abundant. Concordantly, low abundance of non-coding RNAs makes it impossible to achieve the same levels of detection as protein-coding genes, since the presence of ORF and TSS signals enable them to be detected even at low expression levels. Additionally, a sequence homology search can also detect protein-coding transcripts from a phylogenetically related species, although this method is not always sufficient for new transcriptomes, owing to the lack of availability of an annotated relative species[34]. In general, detection of novel splice sites and multiple isoforms through RNA-Seq experiments have been boosted, but they present a challenge in accurate reconstruction of the transcriptome, as alternatively spliced isoforms can map to multiple targets, which can be mitigated by increased read

depth. Furthermore, presence of shared exons between isoforms can lead to ambiguous transcript detection[34;35;41].

The clear connection of functions and sequences in case of mRNAs lack in lncRNAs, and at times lncRNAs are trapped into guilt-by-association when it comes to function prediction[4;61]. Secondary structures for lncRNAs are hard to predict, which leads to difficulties in understanding sequence conservation. Conservation of expression, syntenic location in different genomes can also be gleaned from conserved nucleotide patterns[61]. LncRNAs are also less stable, apart from being less abundant, and many localise to the nucleus, unlike mRNAs, leading to emergence of different observed functions, such as chromatin remodelling. Alternative splicing occurs in some lncRNAs, although in general they follow canonical splicing and possess a lesser degree of polyadenylation[62]. To achieve functional annotation of lncRNAs, it is imperative to determine interacting elements within and without the transcripts, functional importance, and attachment to biological processes[61].

To build an annotation, there are several factors that need to be considered that include the comprehensiveness of the annotation, how many of the transcripts from each locus are compiled, and if all the important physical signals of a gene are represented. Smaller annotations are more complete than larger ones; with a larger swathe of the transcriptome to cover, larger annotations falters sometimes owing to the presence of false positives[4]. Nevertheless, there are issues that affect the quality of all types of annotations, especially for lncRNAs. The 5 ' and 3 ' ends can be incomplete in short-read sequencing and, for example, missing exons can lead to negative effects, particularly for lowly-expressed lncRNAs[63]. Long-read sequencing can fare better in this respect as the reads encompass whole exons. The incompleteness of lncRNA annotation may also be attributed to performing transcriptome reconstruction from adult tissues, while ignoring embryogenesis and developmental stages[4;64]. In RNA capture sequencing, detection of lowly-expressed lncRNA sequences and full-length transcripts can be boosted by using oligonucleotide probes enriching specific targets[65]. Nevertheless, in terms of isoform diversity, the lncRNA transcriptome seemingly has an upper threshold and can be mapped, albeit the process is challenging[4;66].

## 3.2 Annotation catalogues

In this part, a few of the most comprehensive lncRNA annotation catalogues used by researchers will briefly be discussed. The GENCODE consortium produces a tri-monthly update on the human genome annotation with the latest release being GENCODE version 35[5](www.gencodegenes.org). It includes 17,957 lncRNA genes out of 60,656 annotated genes, with protein-coding genes numbering at 19,954. It follows a certain scheme while annotating the lncRNA genes and it has been widely accepted. After several revisions, they are divided into five different categories.

- *Intergenic:* These lncRNA transcripts are located in the intergenic space between two protein-coding genes. Also called lincRNAs, the genes of this biotype are characterised by histone H3K4-K36 chromatin signatures[57;67]. However, they

exhibit similar features as mRNAs: 5' end capped, possess poly(A) tail, undergo splicing, and are transcribed by RNA pol-II. Many of them are highly conserved. LincRNAs are usually nuclear localised; one example would be *LINCRNA-P21*, which recruits the nuclear factor hnRNP-K to promoters mediating p53-dependent transcriptional responses[68].

- *Antisense:* As the name suggests, they belong to the antisense strands of protein-coding genes and are also called asRNAs. They are evolutionarily poorly conserved. These lncRNAs can be further subdivided into two groups: *cis*-NATs (natural antisense transcripts) regulating sense transcript expression and *trans*-NATs regulating non-paired gene expression from other genomic locations. They are more stable than lincRNAs[68]. An example is *BACE1-AS* which is overexpressed in Alzheimer's[69]. *AS-Uchl1* includes inverted short interspersed nuclear element B2, or SINEB2, and is part of a NAT group called SINEUPs can pair to mRNAs and stimulate mRNA translation. SINEUPs could become a potential synthetic reagent in therapy of haploinsufficiencies[70].

- *Bidirectional:* These lncRNAs are produced from the opposite strand of a coding strand and can partially overlap the 5' end of the paired coding gene. They are highly unstable, but bidirectional promoters have been found to show specific epigenetic features and to be located near genes related to cell cycle regulation, for example[68].

- *Intronic:* Complying with the name, these lncRNAs originate from introns of coding genes. They could be produced from pre-mRNA processing. An example would be a lncRNA that is a snoRNA precursor[71]. Functionally, they are suggested to have a positive regulatory effect on coding gene transcription or on its splicing machinery. Intronic genes have also been attributed to be independent of pre-mRNA processing[68].

- *Overlapping sense transcripts:* Opposite of an intronic lncRNA, these genes do not overlap with sense exons, rather surround an exon or coding gene entirely, transcribing in the same sense direction. *SOX-OT* is an example which contains the entire *SOX2* gene in its intron, a pluripotency regulator[72].

Non-canonical splicing of intronic and overlapping sense lncRNAs can lead to the formation of circular lncRNAs (circRNAs). These RNAs have been seen to act as miRNA sponges in the cytoplasm; for example, *CDR1as/ciRS-7* acts as a sponge to miR-7[73]. CircRNAs have been implicated in disease associations as well[74].

Reference Sequence (RefSeq) (www.ncbi.nlm.nih.gov/refseq) is an annotation catalogue similar to GENCODE that incorporates cDNA, EST, and RNA-Seq information. It contains manually curated genes as well as sequences from Illumina, covering multiple species[75]. Its current iteration is release 203 encompassing 105,349 organisms.

The Functional Annotation of the Mammalian Genome CAGE-associated transcriptome (FANTOM-CAT) (fantom.gsc.riken.jp/cat) is an endeavour to map 5' ends of human lncRNAs along with annotating transcription signals and expression data. The effort resulted in 23,887 lncRNA genes[76].

An automated annotation stemming from RNA-Seq data is MiTranscriptome, which at its inception contained 58,648 human lncRNA genes, most of which are expressed in tumour cells. Around 46% of the genes appeared to be annotated in the contemporary annotation catalogues[77]. (The web resource cannot be accessed currently.)

One of the biggest non-coding RNA databases around is called NONCODE[11] (www.noncode.org/index.php). Currently in its sixth iteration, it lists non-coding genes from across 39 species, 16 of them animals, including human, rhesus, chimpanzee, mouse, orangutan, and pig, and the rest plants, such as *A.thaliana*, cucumber, wheat, soybean, and maize. The authors present a curated database of ncRNAs by surveying literature and experiments conducted. The focus of this database is lncRNAs and in its latest version 96,411 lncRNA genes and 173,112 transcripts for human have been compiled.

RNAcentral (rnacentral.org) is an integrative annotation catalogue which amasses non-coding RNA information from numerous databases including not only lncRNAs, but also miRNAs and snoRNAs of many species[78]. It provides stable identifiers for the distinct RNA sequences and functional annotation for some species. It is currently in its $16^{th}$ iteration.

LncRNAWiki (bigd.big.ac.cn/lncrnawiki/index.php/Main_Page) provides researchers with manually curated functional annotation of human lncRNAs besides disease association and putative peptide information. The latest iteration contains 106,063 lncRNA transcripts[79].

LNCipedia (lncipedia.org) is another integrative annotation catalogue that curates human lncRNA information from literature and other databases. It provides coding potential and locus conservation information. In its present iteration (ver 5.2) it comprises 107,039 high-confidence lncRNA transcripts from 49,372 high-confidence genes[80].

## 3.3 Evolution of perception of RNA as a regulator

Throughout the changes that have occurred in all organisms from the beginning of life, molecular biology has revolved and continues to revolve around RNA. Around the time Friedrich Miescher (in 1869 to be exact) discovered DNA, researchers believed that carriers of genetic information were proteins; DNA was called 'nuclein', though, as it was detected in the nucleus[81]. Much later, in the 1940s, DNA was accepted to encode genetic information[82]. That was followed by the establishment of the DNA-RNA-protein network proposed by Crick in 1958 in form of the *central dogma of molecular biology* (Fig. 3.1), which describes how the genetic information encoded in the DNA is converted into functional products called proteins and this message is ferried by RNAs[68;83;84]. This flow of information, however, was posited to be unidirectional, *i. e.* no information could flow from the proteins to the nucleic acids, a view that has mostly changed over recent years.

The RNA that acts as a messenger - hence, messenger RNA (mRNA) - transports the genetic information out of the nucleus into the cytoplasm to sub-cellular components

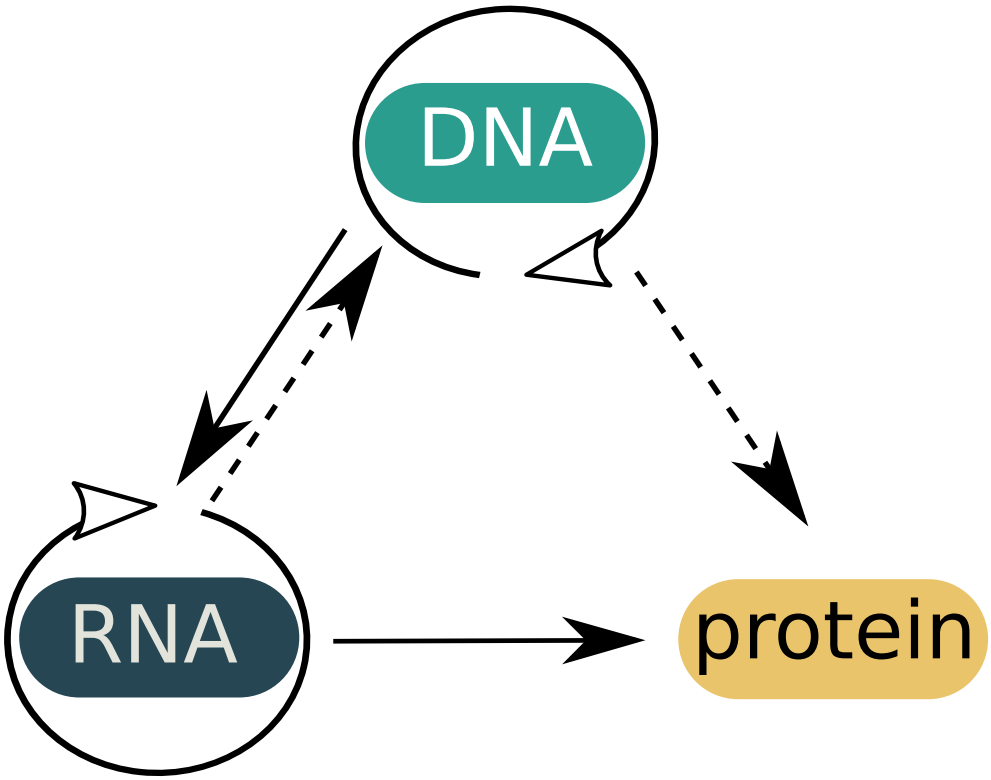



**Figure 3.1:** The central dogma. Genetic information is transported from the DNA to the RNA, which in turn goes to proteins, as was proposed by Crick. The dotted lines represent infeasible information flow directions.

called ribosomes, which in turn participate in protein synthesis as part of the ribonucleoprotein complex (RNP).

Shortly afterwards, amino acids as tri-nucleotides were described to be part of the protein synthesis machinery[85]. Subsequently, mRNA and ribosomal RNA (rRNA) precursors were discovered in the cell, which in turn led to the discovery of small nuclear RNAs (snRNAs), establishing a new hitherto unseen avenue in RNA processing machinery: *splicing*[86;87]. Small nucleolar RNAs (snoRNAs) were also discovered and found to be interacting with the spliceosome and facilitating the processing of rRNAs in the nucleus[88]. Several groups studying the transcription and translation machineries discovered ribozymes, which are crucial in the information flow and are found in both the spliceosome and the ribosome, and are functional in degrading RNA molecules[89]. Evidence suggests that RNA is active and a necessary molecule in a multitude of activities in cell biology. For instance, RNA is required for DNA replication and ribonucleotides are actually precursors of deoxyribonucleotides. There are several small RNA categories which function as regulatory and housekeeping entities, such as small interfering RNAs (siRNAs), Piwi interacting RNAs (piRNAs)[90]. Besides mRNAs that transport the genetic code to the cytoplasm to be translated, rRNAs and tRNAs, among others, are also active in protein synthesis[91]. The first two miRNAs to be discovered were *micF* in the bacteria *Escherichia coli* (*E. coli*)[92] and in the *lin-4* gene of *Caenorhabditis elegans* (*C. elegans*), followed by the discovery of *let-7*[93]. Concurrently, the understanding of the regulatory machinery of RNAs slowly began to increase.

### 3.3.1 Discovery of lncRNAs

The first long non-coding RNAs (lncRNAs) were discovered in the 1980s. Molecular studies on genomic imprinting led to the discovery of two genes, *IGF2R* and *H19*, who formed the *IGF2R/H19* cluster. *IGF2R* was found to be paternally expressed, while *H19* was maternally expressed, and both the genes were classified as protein-coding genes. *H19* showed mRNA-centric features of RNA polymerase-II (pol-II) transcription, polyadenylation at the 3*'* end, splicing, and localisation in the cytoplasm, although lack of translation was noticed despite the presence of small open reading frames (ORF), but no large ones. *H19* was even found to be highly conserved, a signature of protein-coding genes[94;95] and was only presumed that this RNA was involved in embryonic development. Shortly afterwards, *XIST* was discovered, which was found to be playing a very important role and is investigated until this day. The X-inactivation center (*Xic*) locus, which is home to the phenomenon of one of the female X chromosomes inactivation, was found to generate the lncRNA *XIST*, that is localised in the nucleus and triggers *cis* gene silencing, whereby it silences the surplus X chromosome in embryonic cells. Already *XIST* was found to recruit polycomb repressive complexes 1 and 2 (PRC1, PRC2) to carry out chromosomal repression, although it was not the only player[96]. Slowly, other genes were discovered in the same locus. The lncRNA *TSIX* was observed to be the overlapping antisense of *XIST* in mouse, but in human it overlaps only the 3*'* end of *XIST*. The *Xic* locus is longer than 1 megabases (Mb) and produces several protein-coding and non-coding genes[97]. The evidence of non-coding RNAs, other than protein-coding RNAs, was already changing the way researchers viewed genomic space of eukaryotes[92]. The discovery of transcripts that act like RNAs but did not have sufficiently large ORFs for translation opened up a new avenue in molecular biology.

When the Human Genome Project burst into the scene, the race to sequence the human genome gathered further momentum. It reported around 19,000 protein-coding genes to be present in the human genome and further inferred that about 1.2% of the genome contained protein coding exons and the rest were intronic and intergenic non-coding DNA[98]. A study on human chromosomes 21 and 22 employing tiling arrays with oligonucleotide probes detected cytosolic polyadenylated transcripts originating from the non-coding region of the genome to be around 90%[99]. This was confirmed by the FANTOM consortium studying mammalian transcriptomes, adding that perhaps two-thirds of the mammalian genome did not code for proteins[100]. The Encyclopedia of DNA Elements, or ENCODE consortium, studied 147 cell lines discovering far reaching results in the human genome annotation process[101]. Using sequencing techniques, histone modifications, DNAse I hypersensitive sites, and transcription factor binding sites were analysed to primarily report that the human genome is comprehensively transcribed (around 93%), although since then it is found that the transcribed area is much less, around 80%[24]. It was noted that many lncRNAs overlap with protein-coding genes in both directions have at least one primary transcript, with about 54% of the transcripts mapped outside coding genes. The GENCODE consortium[5;102] was conceptualised within the ENCODE framework and focused on protein-coding genes in human transcriptome, later scaled up to include non-coding genes. The emergence of evidence of a non-coding transcriptome was in contrast with the tenet of cell biology, the central dogma (Fig. 3.1), especially since everything outside the coding regions was considered to be transcriptional noise[103].

### 3.3.2 Molecular structure and gene regulatory mechanisms

New lncRNA genes are constantly being discovered and many of them have been shown to have regulatory roles in human and other species[1;104]. For example, *KCNQ1OT1* is involved in genomic imprinting[105] and *COLDAIR* is functional in development[106]. Evidence of regulation of the epigenetic landscape and gene expression, and histone modification by lncRNAs have been shown[67;107] and their involvement in pluripotency and *p53* response pathways were also some aspects studied[57;108]. In the following part of the chapter, different types of lncRNAs and their regulatory roles will be described.

#### Molecular structure

Non-coding genes, in general, were described to be produced because a myriad of factors, such as DNA-based duplications of existing sequences, transposable elements or non-coding DNA exaptation, metamorphosis of protein-coding genes due to the loss of coding potential[109]. For instance, the lncRNA *PTENP1* is transcribed from a pseudogene that is found be a product of metamorphosis of a protein-coding gene; translation possibilities have disappeared because of disruptions in the duplicated ORF. The pseudogene *LNX3* produces *XIST*, which contained frame-shifting mutations in its ORF. This implies that it is not always necessary to have a coding homologue. Transposable elements (TE) occupy 45% of the human genome and are also responsible for the origin of lncRNAs. The mRNA-like features of lncRNAs such as poly(A) tail and RNA binding sites, along with transcription start sites (TSS), splicing machinery and RNA editing are considered to be the direct influence of TEs[110].

LncRNAs have been discovered in mammals, plants, fungi, and viruses, and are generally accepted to have a lower bound of 200 nucleotides (nt), but can grow to more than 10 kilobases (kb) (Fig. 3.2). Most are weakly polyadenylated and are located in the vicinity of coding promoters, if not overlapping them. Similar to mRNAs, they might have a 5' terminal methylguanosine ($m^7G$) cap. They up-regulate nearby genes in *cis*[111]. They do not code for proteins, however, some lncRNAs are translated to form small peptides, for example, *DWORF*, which forms a 34 amino acids functional peptide[112;113]. Cytosolic lncRNAs create an association with mono and polyribosomal complexes[114]. There are mRNA-like intergenic sequences that lie in the so-called intergenic space between two genes and antisense transcripts of protein-coding genes and are known as long intergenic non-coding RNAs (lincRNAs)[1]. Like mRNAs and miRNAs, most eukaryotic lncRNAs are transcribed by RNA pol-II cleaved by Rnase P to generate a mature 3' end of a U-A-U triple helix structure[1], with a low accumulation of pol-II around the promoter in lncRNAs, unlike mRNAs[111;115]. About one-third of ancient lncRNA promoters were reported to have homeobox transcription factor (mainly *OCT4*) binding sites, twice as much as protein coding genes. *SUZ12* (a member of *PRC2*) was also found to bind preferably to lncRNA promoters; *SUZ12* and *OCT4* being regulators of pluripotency[116]. RNA pol-III transcribes human nueroblastoma associated *NDM29*, although it is less than 200 nt long[117]. Some plant lncRNA genes are transcribed by RNA pol-IV and pol-V[118]. A 5' end cap is observed in several lncRNAs, but there can be a poly(A) tail at their 3' ends, which can be found in bimorphic lncRNAs like *NEAT1* and *MALAT1*. It has also been reported that polyadenylated lncRNAs

can possess higher stability[24;68]. LncRNA transcription takes place with the help of a promoter sequence, pre-initiation complex, transcription elongation complex and other factors[119]. The lncRNAs acting as precursors to smaller RNAs are processed by specific endonucleases[120].

Similar to mRNAs, lncRNAs possess multiple exons and an affinity towards two-exon transcripts, albeit fewer and longer, and undergo alternative splicing to form different isoforms[56;116]. A study by Cabili et al. showed intergenic RNAs to feature 2.9 exons to 10.7 in their mRNA counterparts, and the length of the transcripts were also categorically smaller on average (1 kb to 3 kb). LincRNA genes that were alternatively spliced could lie within close proximity of coding regions or $>$ 3 Mb away[121]. LncRNA genes tend to have multiple isoforms, but it has been seen that there is only one major isoform with a high expression, when the alternative isoforms do not have similar expression levels. In contrast, most protein-coding genes produce at least two different dominant isoforms[24]. Nuclear localised single exon transcipts are unstable. LncRNA molecules have been known to fold, *i. e.* possess complex secondary structures, which enables them to interact with proteins, stabilise, and localise[68]. In human, secondary structure is conserved in about 14% of the genome and it was reported that at least one exon was overlapping with $>$ 90% of the structured segment in one of three of lncRNAs, more than mRNAs, whose share was one-fourth (and one-sixth in strictly coding exons)[122]. Interestingly, Khalil et al.[67] reported an average of 4 exons for every K4-K36 domain.

Cellular lncRNA abundance is controlled by nuclear exosomes in the nucleus and cytoplasmic exosomes, cytoplasmic *XRN1*, nonsense-mediated decay and RNAi pathway. *NEAT1* and *MIAT* are nucleus localised lncRNAs, while *DANCR* and *OIP5-AS1* can be found localised in the cytoplasm, whereas *TUG1* and *HOTAIR* can be found in both nucleus and cytoplasm.[123;124]. Chromatin-enriched and chromatin-associated lncRNAs are suggested to be involved in guiding of chromatin modifications, assembly of RNP complexes, and regulation of protein activity[68]. Djebali et al.[24] observed that chromatin modification by RNAs is related to splicing regulation; exons being spliced are enriched in chromatin marks. LncRNAs are less abundant than mRNAs in the cell and significantly less evolutionary conserved than other RNAs[1]. They are found to be more nuclear localised than mRNAs due to poor splicing machinery and polyadenylation and are susceptible to degradation by chromatin exosomes, also having close associations with nuclear proteins because of the presence of *cis* elements[125]. A specific motif (BORG) was also discovered which primarily occur in lncRNAs localised in the nucleus[126]. Primate specific short interspersed nuclear elements (SINEs), a C-rich sequence from *Alu* elements (a form of transposable elements), interact with the nuclear matrix protein HNRNPK and this act becomes influential in nuclear retention of lncRNAs[1].

LncRNAs that originate from pseudogenes have been found to serve as miRNA sponges, indicating that they can regulate expression through RNA-RNA pairing; for example, *PTENP1* is a well-known miRNA sponge functional in cancer[127]. LncRNAs derived from ultra-conserved genomic regions between human, mouse and rat are suggested to act as decoys; an example is *EVF2* that induces chromatin remodelling by interacting with the transcription activator DLX1 and represses transcription. *EVF2* also represses *BRG1*'s ATPase activity[128]. Non-coding telomeric repeat containing

RNAs, *TERRA*, are transcribed from telomeres, the nucleoprotein structures at chromosome terminals. LncRNAs derived from this region are subTERRA and are thought to induce telomere shortening in cells where there is no telomerase directed repair[68;129]. Promoter and pre-rRNA antisense (PAPAS) lncRNAs are transcribed from the antisense regions of ribosomal RNA that have been reported to be active in histone modification. *FOXC1e* and *NRIP1e* are examples of enhancer derived lncRNAs that are characterised by 3' end transcript cleavage, hence they may not possess poly(A) tails and are highly unstable[68;130].

Promoter-associated lncRNAs are located at the promoter region of a gene and may overlap the 5' end. They are bidirectionally transcribed and are known as unstable promoter upstream transcripts (PROMPT) and upstream antisense RNAs (uaRNAs). PROMPTs, although rapidly degraded assisted by polyadenylation, and uaRNAs, banking on a splicing competent intron, secure directional promoter execution, ensuring RNA pol-II act towards the nearby coding gene. These groups are still being studied to find a coherent function[131].

There are a few areas in the human genome called human accelerated regions (HAR) that diverge faster than in other species and have been identified to produce several lncRNA genes. *HAR1F* was found to be transcribed from *HAR* and observed to be expressed in cortical brain development[68]. It is also speculated that *HAR* enhancer elements could be involved in autism. Some lncRNAs (antisense to coding genes) hold sequence complementarity to sense paired mRNAs, hence through RNA-RNA pairing can target asRNA regulatory activity with *BACE1-AS* being an example, which is overexpressed in Alzheimer's disease, stabilizing *BACE1* mRNA to induce *BACE1*-encoded beta secretase increased expression[69].

### Chromatin remodelling

LncRNAs as regulatory agents can manifest their roles in several ways, however, the most fascinating roles are their regulatory roles, such as scaffolds, guides, and facilitators of ribosomal repression or activation (Fig. 3.3). They can also serve as sponges to miRNAs and as precursors to smaller ncRNA genes, the latter being a strong focus of this thesis. Concomitantly, they act as scaffolds in their tertiary structured form to organise RNP complexes in the nucleus and evidence suggests that they act either in *cis* or *trans* to their transcription sites and exert epigenetic control on gene expression[133]. This is one of the most important regulatory mechanisms that lncRNAs perform, since chromatin modification and reshaping the epigenetic landscape have extended effects including transcription, RNA processing, and DNA repair[134;135]. A well-studied lncRNA, *HOTAIR*, interacts on one hand with PRC2 and Lsd1/REST/coREST complexes in the nucleus affecting histone modifications and gene silencing, and on the other hand with *Dzip3* and *Mex3b* assisting in proteolysis of *Ataxin-1* and *Snurportin-1* in senescent cells[135]. The nuclear localised *NEAT1* assembles RBPs and transcription factors in paraspeckles, where numerous proteins are localised[1;136]. *MALAT1* should also be mentioned in connection to paraspeckles, since it also localises there besides in the cytoplasm and may also be involved in transcriptional regulation[53;137]. It interacts with SR proteins involved in RNA splicing and the dysregulation of *MALAT1* expression

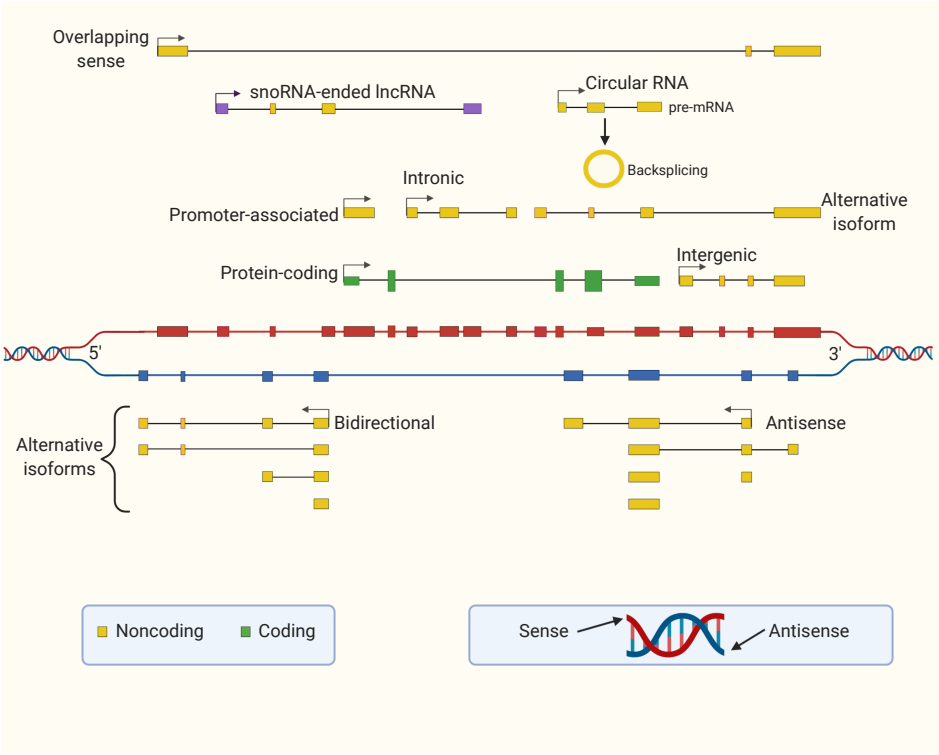



**Figure 3.2:** Diverse lncRNAs. Some examples of lncRNA biotypes. Depending upon their genomic location with respect to protein-coding genes, there can be different types of lncRNA transcripts. They can be organised bidirectionally or antisense to protein-coding genes. Intronic lncRNAs originate from the introns, whereas intergenic lncRNAs lie in-between protein-coding genes. Overlapping sense lncRNAs overlap the whole sense strand of a protein-coding gene. Some exons of pre-mRNAs are back-spliced to form circular RNAs (circRNAs). SnoRNPs at fringes of lncRNA intronic sequences lead to the formation of snoRNA-ended lncRNA (for example, *SLERT*). Alternative isoforms of lncRNAs are results of differential splicing event. The figure is drawn after and inspired by Mercer and Mattick [132] and Yao et al. [1].

leads to pre-mRNAs in cancer cells being not targeted by SR proteins [137]. *MALAT1* also interacts with alternatively spliced pre-mRNAs and it has been suggested that it may enhance protein-protein, protein-RNA and protein-DNA interactions at the paraspeckle level [1].

*MEG3* acts as a guide lncRNA, the group of lncRNAs guiding RNP complexes to chromatin loci, and guides the EZH2 subunit of PRC2 to TGF$\beta$-regulated genes [138]. *GAS5* acts as a decoy by interacting with a glucocorticoid receptor (GR) and preventing it to bind to its GR binding element (GRE), thereby repressing GR-regulated genes [139]. Certain lncRNAs share partial sequence similarity to coding transcripts, thereby competing for miRNA binding sites that induce post-transcriptional regulation, which in turn can regulate mRNA stability in the cytoplasm, and are called competing endogenous RNAs, or ceRNAs [68;140;141]. The well-researched lncRNA *HULC* acts as a sponge to miR-372 and leads to translational derepression of *PRKACB*, which subsequently activates *CREB*, up-regulating *HULC* in liver cancer cells. *LINC-MD1* acts as a sponge to two miRNAs, miR-133 and miR-155, and regulates transcriptional levels of muscle-specific genes [142]. However, this theory of sponges may not be easily described due to complicated results of a study showing how a mammalian brain cell circular lncRNA *CDR1AS*, which has a binding site to miR-7, would be affected, if other ncRNAs were

also present[1]. *H19*, hosting miR-675-3p and miR-675-5p, is a precursor lncRNA, which is active in post-transcriptional regulations of the anti-differentiation transcription factor Smad[143]. *MALAT1* hosts a cytoplasmic mascRNA processed by RNase P and Z cleavage machinery[68], while *NRON* regulates movement of the transcription factor NFAT, which is transported to the nucleus from the cytoplasm in order to activate target genes by interacting with importin-$\beta$ proteins[132].

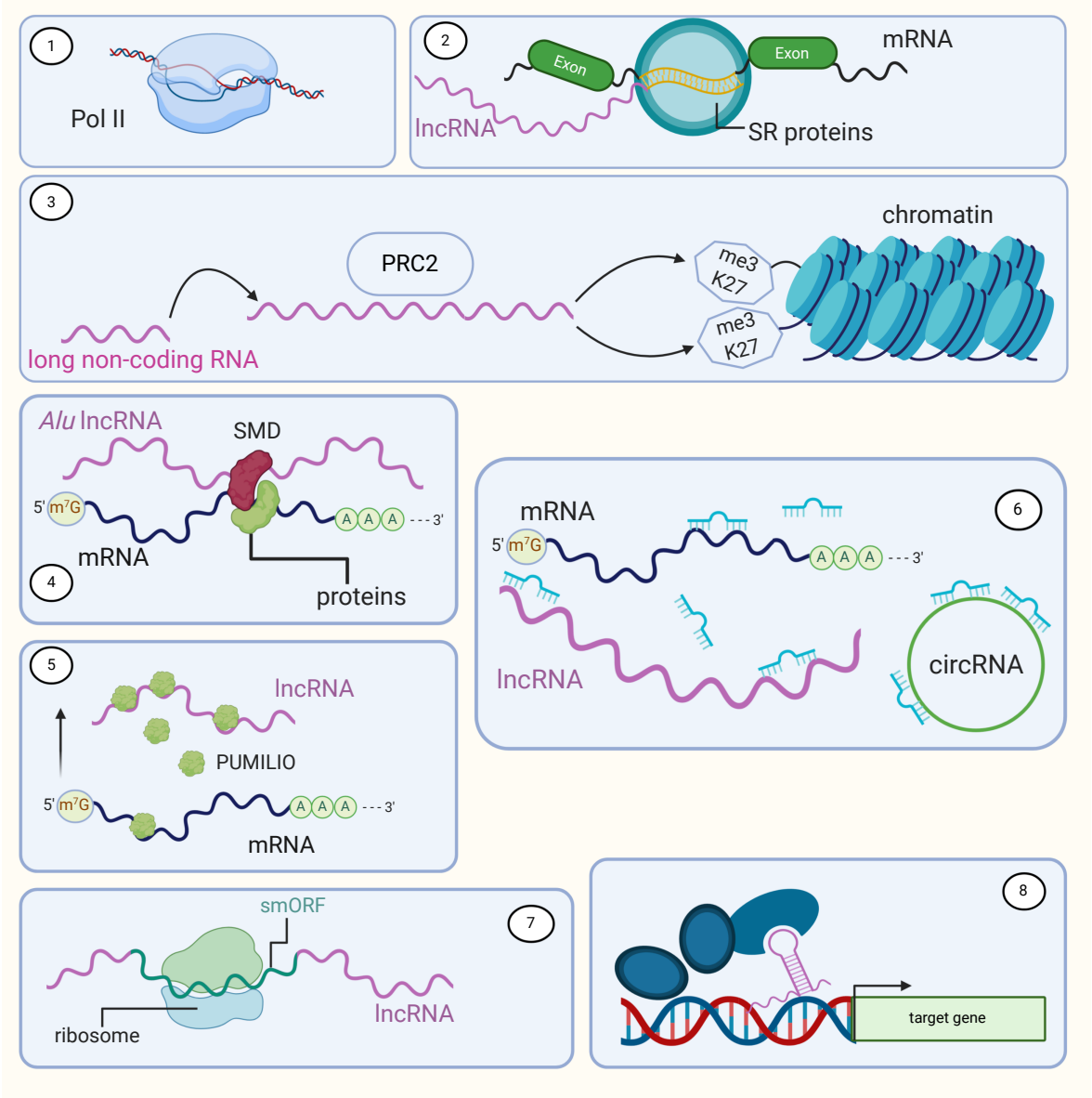



**Figure 3.3:** Regulatory roles of lncRNAs. LncRNAs regulate gene expression and also participate in chromatin remodelling and post-transcriptional regulatory activities. **(1)** LncRNAs are primarily transcribed by RNA pol-II. **(2)** Pre-mRNA splicing is impacted by the lncRNA *MALAT1* when it interacts with SR proteins and alters their phosphorylation. **(3)** LncRNAs like *HOTAIR* recruit chromatin modifying complexes (like PRC2) to regulate gene expression of protein-coding genes. **(4)** *Alu*-containing lncRNAs modulate mRNA stability by recruiting STAU1 to induce STAU1-mediated mRNA decay(SMD). **(5)** *NORAD* sequesters *PUMILIO1/2* from mRNAs, thereby stabilising *PUMILIO1/2*-targeted mRNAs. **(6)** LncRNAs compete for miRNA targetting. The miRNA miR-7 targets circRNA *CDR1AS*, thereby resulting in lack of repression of its target genes. **(7)** LncRNAs can be translated, if there is a small ORF present, forming peptides (for example, *DWORF*). **(8)** LncRNAs like *XIST* can act as scaffolds. This figure is drawn after Mercer et. al[107], Kapusta and Feschotte[122], Yao et al.[1], and Robinson et al.[3].

Certain lncRNAs act as decoys preventing chromatin modifiers to bind to specific DNA loci. One example is *GAS5*, which is already mentioned. Myosine heavy chain-associated RNA transcripts (*MHRT*), group of lncRNAs associates with RNA helicase domain, forcing *BRG1* not to target the genomic loci[1].

R-loops are triple stranded nucleic acid structures with RNA hybridised to duplex DNA. Certain antisense lncRNAs can form R-loops to recruit transcription factors to promoter regions in order to regulate sense mRNA transcription, either in *cis* or *trans*[1]. *VIM-AS1*, transcribed from human *vimentin* (VIM), manipulates local chromatin decondesation to induce NG-$\kappa$B binding to the VIM promoter by forming an R-loop at VIM TSS[144].

### Post-transcriptional regulation

LncRNAs can influence transcription of other genes, even if they are not directly involved at the transcript level. This phenomenon has also been reported true for protein-coding genes. *SFMBT2* gene expression was impaired when *BLUSTR* was not allowed to be transcribed through promoter deletion, polyadenylation, and 5'-end splice site mutation. The lncRNA *AIR* overlaps with the promoter region of *IGF2R*, its sense gene, and prevents pol-II from being recruited to the chromatin; consequently *IGF2R* is silenced[145]. The transcribed Upperhand (*UPH*) becomes an enhancer for local gene regulation in its parent locus of cardiac TF Hand2. Plasmacytoma Variant Translocation 1 (*PVT1*) promotes MYC degradation by interacting interfering with the protein's Thr58 phosphorylation. It is to be noted that MYC lies 55kb upstream of *PVT1*[1].

### mRNA stabilisation

Stability of mRNAs can be regulated by lncRNAs containing *Alu* elements; the *Alu* elements can form double-stranded DNA molecules through reverse transcriptase after being transcribed into an mRNA and can contribute to formation of new genes. These lncRNAs can form double-stranded RNAs (dsRNAs) with the target mRNAs, thereby down-regulating them, since the dsRNAs contain binding sites to Staufen 1 (STAU1) resulting in Staufen 1-mediated mRNA decay (SMD) in *trans*[146]. In contrast, lncRNAs like *NORAD* (noncoding RNA activated by DNA damage) can act as decoys to RBPs like PUM1/2 that degrade mRNAs[147]. PUM1/2 bind to mRNAs at 3'-end to facilitate deadenylation and decapping, leading to increased turnover and decreased translation. The binding of PUM1/2 to *NORAD* is thought to be a necessary operation as knockdown of *NORAD* increased choromosomal instability, possibly due to degradation of mRNAs encoding proteins for stability[148]. LncRNAs are not translated, despite the presence of ribosome-associated lncRNAs, however, they can regulate mRNA translation. The mRNA *Uchl1* is promoted towards its translation by its antisense transcribed lncRNA *AS-UCHL1*[70].

### Case in point: XIST and HOTAIR

As already mentioned, lncRNAs play an important role in regulating chromosome architecture[1]. X chromosome inactivation (XCI) has been one of the areas extensively researched including the X-inactive-specific transcript, or *XIST*. XCI occurring in female mammals takes place during early embryonic development, when *XIST* silences the whole future X inactive chromosome (called Xi) from which it is transcribed by performing chromatin remodelling[149;150]. It was initially observed that *XIST* localised to the Barr body and coated the inactive X[151]. The lamin B receptor (LBR) is associated with *XIST* and whose depletion leads to disruption in XCI suggesting that *XIST* might play a significant role in chromatin modification by recruiting Xi to the nuclear lamina[152]. A subsequent role of *XIST* shows it to interact with PCGF3/5-PRC1 complex and the ensuing ubiquitylation of histone H2A lysine 119 (H2AK119u1) facilitates the recruitment of PRC2 triggering H3K27me3 dependent chromatin repression in Xi[1]. Later, *XIST* interacts with silencing mediator for retinoid and thyroid hormone receptor/histone deacetylase 1-associated repressor protein (SHARP or SPEN), directing histone deacytelation and transcriptional repression[153–155]. Deleting the A-repeat region on *XIST* ablates its silencing role but does not change its Xi localisation, which also proves that *XIST* is crucial in XCI[155]. This is how *XIST* performs a vital role in chromatin remodelling. Long interspersed nuclear elements (LINEs) and some highly repetitive RNAs are shown to be associated with euchromatin and could play a role in chromatin opening. Another lncRNA transcribed from the X chromosome is functional intergenic repeating RNA element, or *FIRRE*, does not participate in XCI, rather binds to hnRNPU and is present in five different autosomal chromosomal loci, the combined locus being 5Mb long. It is theorised that *FIRRE* serves as a scaffold to modulate interchromosomal interactions, since depletion of either the RNA or the protein leads to *FIRRE* not being accumulated at the loci. The colorectal cancer associated transcript 1, long isoform, or *CCAT1-L*, accumulates at its transcription site to promote the transcription and has an oncogenic effect on the MYC locus[1]. *CCAT* promotes long-range chromatin looping[76].
*HOTAIR*, short for Hox antisense intergenic RNA, transcribed from HoxC locus, recruits PRC2 to the HoxD locus and interacts with H3K37me3 and suppresses HoxD expression in *trans*[106]. It has also been noted that artficial tethering of *HOTAIR* to a luciferase reporter locus led to PRC2 independent repression in breast cancer cells[156]. However, *in vitro* bindings and RNA immunoprecipation experiments have resulted in false positive interactions, which insinuates that it is still not a well-understood process[1].

### *cis* or *trans* regulatory function

While defining functional roles of lncRNAs, it is essential to see whether they act as *cis* regulatory elements (at sites near their transcription sites) or in *trans* (at sites away from their transcription sites). Several lncRNAs first observed or described to have functions based upon their RNAi pathway were discovered to have functions quite disconnected to their products, occurring in *cis*[145;155;157]. An example would be *LINCRNA-P21* that was previously described to be acting in *trans* based upon its RNAi pathway[108], however further experiments suggested that the phenotype could be understood by the *cis* regulatory function of the DNA element and the RNA product could be

ignored[111;155;158]. The lncRNA *Oct4P4* acts in *trans* to interact with SUV39H1 HMTase and removes SUV39H1 and H3K3Me3 marks at Oct4 promoter. *BRAVEHEART* interacts with the PRC2 complex and regulates MesP1 in *trans*. *HOTTIP* interacts with WDR5 in the HOXA locus in *cis* to drive gene expression. *COOLAIR* silences the FLC locus by binding to it in *cis* and reducing H3K36me3 levels[159]. The mechanism behind localisation of lncRNAs can provide the clue to their functions, since they are functional molecules and must remain close to their regulatory sites[155]. Tuck et al.[111] predicted by knocking down lncRNA loci to observe regulatory effects on nearby genes that a third of lncRNAs act in *cis* and most lncRNAs terminate transcription early and degrade, consistent with the suggestions that these RNA genes do not require their product to function, also given local enhancer-like properties of some lncRNA loci. LncRNA transcription processes were also found to be bidirectional from their TSSs, although susceptible to quick termination (50-300 nt from TSS), in contrast to mRNAs which favoured forward and more persistent transcription. Since lncRNAs are co-expressed with their neighbouring genes over both short and large genomic distance[111], they argued this could be the reason for *cis* regulatory effects to prevail notwithstanding the possibilities of co-regulation of nearby gene pairs. They also wanted to observe global effects of transcription in lncRNA gene deletion cell lines and in half the loci they experimented upon, certain evidence of functions in *trans* was seen. However, further experiments are required to establish which mode of function lncRNAs prefer[111].

### Evolutionary properties

An RNA localised in the cytoplasm would be involved in post-transcriptional regulation and translation of mRNAs, and not in chromatin modification, whereas a nuclear RNA would not be translated. Interaction with proteins or bindings with other enzymes could also explain their sub-cellular localisation, whereby they are transferred to a different location than their transcription origins. Fluorescence in situ hybridisation (FISH) showed that *AIR* and *KCNQ1OT1* are found on the same allele near their transcription sites and they co-localised with repressive chromatin modifiers, suggesting a *cis* regulatory function[155]. *MALAT1* and *NEAT1* were shown to localise to transcribed DNA loci throughout the nucleus[53;155]. A lowly expressed lncRNA *HOTTIP* is found near its transcription locus[155], a direct contrast to *MALAT1*. However, in spite of documented observations of lncRNA localisation, a concrete relation between that and lncRNA functions still need to be established.

Cabili et.al.[104] showed that lncRNAs exhibited nucleus-preferred localisation and more than 95% of the 61 lncRNAs in three cell types had higher nuclear fraction than mRNAs. They noted that a particular sub-cellular localisation pattern was quite independent of gene expression correlation of two neighbouring genes, one of them being an lncRNA. The authors also found out that cell-to-cell heterogeneity of lncRNAs was at a similar level of mRNAs. Matching pairs of divergent lncRNAs and mRNAs were not always co-regulated. A previous study conducted across 24 human tissues and cell lines discovered that the intergenic transcripts were highly tissue specific, unlike protein-coding genes, consistent with Mercer et al.[107]. They also reported categories of orthologous transcripts of human lincRNAs in other vertebrates, highly conserved

lincRNAs with low coding potential, and lincRNAs in disease associated regions[121]. Their results also showed that protein-coding genes close to lincRNAs were associated with developmental and transcriptional regulation and the results were coherent with previous studies[57;121]. The lincRNA genes were co-expressed with their protein-coding neighbours, but not more correlated to a lncRNA - protein-coding gene pair than a protein-coding - protein-coding gene pair, but this result also threw some light on the *cis* or *trans* behaviour in lncRNA functions. That lincRNAs exhibited high tissue specificity was also reiterated by Washietl et al.[160], who conducted a study to characterise tissue specificity, splicing patterns, and expression levels across nine tissues in six mammals, and they suggested that it was selectively maintained across evolution. Both the studies showed evidence of specificity in brain and testes, sites of highly expressed lincRNAs; confirmed later in a separate study, too[76].

Splicing patterns of lincRNAs were also observed to be highly divergent and not essential to their functions, however, intra-species conservation levels were similar to protein-coding genes. Concurrently, expression conservation dropped towards evolutionary distant species, quicker than sequence conservation, in contrast to mRNAs, where the levels were constant, suggesting that lincRNAs had a higher turnover than mRNAs, even in closely related species. It is worth noting that, gene expression levels across long non-coding RNAs are measured to be less than protein-coding genes[24;116;160]. Concurrently, the population of younger lncRNAs is more enriched than the ancient ones in lncRNA evolution, who have low levels of exonic sequence conservation across evolution[116]. Contrastingly, ancient lncRNAs (>90 million years old) have long-term exonic sequence conservation of higher levels[116]. There seemed to be a regulatory constraint acting in case of lincRNA promoters, similar to mRNAs, and the promoter sequence was reported to be conserved, along with the binding sites in promoter regions[57;116;160;161]. Regulatory conservation was also found to be consistent in lincRNAs, similar to protein-coding genes, despite the former having lower sequence conservation. A low exon conservation was also reported for lincRNAs[24]. Derrien et al. noted that fewer lncRNA transcripts included one of the six most common poly(A) motifs than protein-coding transcripts[56], explaining the characteristic feature of poor polyadenylation in most lncRNA transcripts. A later work also reported that median half life of lncRNAs is lower than mRNAs based upon their expression. The low expression levels of lncRNAs could be defined by exosome-mediated decay of mature lncRNAs. In contrast to mRNAs, lncRNAs were observed to be transcribed into shorter transcripts in the vicinity of the promoter and into fewer full-length transcripts[111].

### RNA secondary structure

The secondary structure of an RNA refers to the formation of structures created when a single-stranded RNA molecule internally folds on itself, forming double helical and stem-loop structures[162]. The base pairing is driven by hydrogen bonding between the nucleotides, where purines normally pair with pyrimidines - AT(U) and GC, with the former having two hydrogen bonds and the latter three and is known as Watson-Crick base pairing. Besides Watson-Crick base pairing, wobble base pairing occurs when a GU pair is formed, and Hoogsteen base pairing, which shows an alternate geometric layout of the AT(U) and GC base pairs[163]. The non-canonical pairing mechanisms can be

due to non-covalent interactions of secondary structures in immediate neighbourhoods of each other, which in turn form complex tertiary structures that are energetically less favoured[162]. The stem-loop structure an RNA molecule is observed when the stem is formed by complementary base pairing, whereas the loop (the hairpin) constitutes unpaired bases[164]. Each secondary structure is thermodynamically active - the folding energy required is defined by the energy required to undo a base pair and to destabilise the entropic effects of the hairpin loops[162]. The minimum free energy of an RNA sequence is given by the secondary structure that is most stable and is widely used to predict the secondary structure of an RNA[165].

Secondary structures can shed light on the functions of a lncRNA, as the base pairs (sequence structure) have been found to be conserved in homologues[61;166]. For example, lncRNAs have already been found to be active in chromatin remodelling by forming scaffolds and interacting with PRC2[122]. Additionally, they interact with DNA to form a triplex; for example, the lncRNA *MEG3* promotes fibrosis by interacting with PRC2 and forming RNA-DNA triplex in the presence of GA-rich sequence binding sites[167]. Although determination of lncRNA structure is challenging, owing to the length of the molecules as well as contrasting regions, where there are regions with well-defined base pairing, while there are others with no base pairing and multiple structures, a few lncRNAs have been closely observed to have well-defined structures[167;168]. The lncRNA *XIST* participates in epigenetic modifications with detected secondary structures[169;170]. The human steroid receptor RNA activator (SRA) possesses several secondary structures and is found in both human and mouse. It produces both an mRNA and a lncRNA and has been evolutionary stable, while stabilising the RNA structural core[167;171]. *HOTAIR* is an enormous lncRNA (12,651 bases), which plays an important role in cardiovascular system, has been observed to have evolutionary conserved sequence elements as well as secondary structures[172]. *BRAVEHEART* transforms into secondary structures consisting of numerous helices and loops. It has been detected in cardiovascular lineage regulation. Moreover, it contains a AGIL loop (5' asymmetric G-rich internal loop) which is essential in mouse embryonic stem cell differentiation[167;173].

There are numerous computational tools for RNA secondary structure prediction, such as[162;174;175]. If those can be utilised in conjunction with experimental determination of lncRNA secondary structure, functional domains of lncRNAs can be exposed.

## 3.4 Disease association

LncRNAs have been implicated in various regulatory roles within a cell, which led them to be performing very important biological functions, like X inactivation. Concurrently, they have also been associated with several diseases. They have been especially found to have roles in several types of cancer. Differential expression of lncRNA isoforms in tumour in comparison with normal tissues led to the observation of lncRNA-cancer association[176;177]. LncRNA genes like *H19*, *MALAT1*, and *PCA3* are over-expressed in multiple cancer cells. *H19* has been found to be associated with the likes of bladder cancer[178], gastric cancer[179], and esophageal cancer[180]. Both *H19* and *MALAT1* have been implicated in colorectal cancer[181;182] and glioma[183]. *MALAT1* is primarily over-

expressed in lung cancer[184], but also active in hepatocellular carcinoma[137], and prostate cancer[185]. PCA3 is found over-expressed in prostate cancer[186]. NF-$\kappa$B interacting lncRNA (*NKILA*) interferes with phosphorylation of I$\kappa$B, which results in NF-$\kappa$B activation and breast cancer metastasis suppression[1]. A deregulated *HOTAIR* gene recruits PRC2 to the promoter regions of tumour suppressor genes, leading to their transcriptional repression and chromatin condensation[177]. *HOTAIR* has been implicated in several types of cancer including esophageal, lung, breast, and pancreatic cancer[1;177]. *ANRIL* behaves in a similar fashion and has been observed to be up-regulated in lung cancer, hepatocellular cancer, and bladder cancer[177;187]. lncRNA *SChLAP1* binds to tumour suppressor complex SWI/SNF and impairs its function, promoting tumour cell division and metastasis. *SChLAP1* is over-expressed in prostate cancer[188]. A similar function was observed in NEAT1, which is over-expressed in oral cancer and glioma, among others[177;189].

Besides various types of cancer, lncRNAs have been found to be dysregulated in other diseases. For example, BACE1-AS is overexpressed in Alzheimer's disease[190]. *NEAT1* and H19 have been implicated in autoimmune diseases and diabetes[191;192]. *SNHG1* has been associated with Parkinson's disease[193], whereas *TapSAKI* and *PVT1* are found in diabetes mellitus and other kidney related diseases[194]. In the Appendix Table A.1, a *non-exhaustive* overview of disease associated lncRNAs can be found.

Databases such as LncRNADisease[195] catalogue lncRNA-disease associations besides regulatory relationships between lncRNAs, miRNAs, and mRNAs.

# 4

# Machine Learning

Computer science has helped humans to overcome seemingly difficult and time-consuming tasks mostly through automation. Humans still have had to anticipate the different parameters and foresee the outputs to be able to efficiently program the codes for the tasks. Now that human intelligence is becoming replaced with **artificial intelligence**, or AI in short, which is a branch of computer science dedicated to *smarter* performing of the same tasks, without requiring human intervention. The machines are designed to be ultimately capable of simulating human behaviour. The applications of AI can be found everywhere today, for instance in search engines, translators using natural language processing, disease prediction, computer vision among many others.

## 4.1 A very short timeline of intelligent objects

The concept of intelligent objects has been around for a long time. History tells us that mankind has been engrossed with inanimate objects with human-level intelligence designed to autonomously perform tasks since ancient times. Beginning with Hephaestus, Daedalus, ancient Egyptian engineers who built statues of gods, to thinkers like Aristotle and René Descartes, although separated by time, who tried to describe the significance of human thought processes as symbols, all of them contributed somehow to the foundation of AI concepts. In popular culture, authors like Mary Shelley to Isaac Asimov to Michael Crichton and numerous films and TV programmes have delved deep into parallel civilizations where automatons with near-human or super-human intelligence are a reality. The question that everyone tried to address was: what underlies our thought processes, our intelligence? The forebearers of modern computer are the programmable machines developed by mathematician Charles Babbage in the early nineteenth century. The series of designs that Babbage developed, named the Analytical Engine, had the potential to solve general purpose computational problems

using punched cards - the first by a mechanised device. Lady Ada Lovelace developed an algorithm to solve for Bernoulli numbers using the machines. About a century later, John von Neumann created an architecture of a computer which could store the program and the processed data in its memory. British mathematician Alan Turing went even further by designing a method by which a computer could devise its own answers to a series of questions with incredible accuracy, so that the answers would seem that they originated from a human being. This method is known as Turing's Test and is widely used as a yardstick for AI applications - if the program can pass the test, it is strong.

*Turing Test* or *Imitation Game*, proposed by the British mathematician Alan Turing in 1950, is a method by which an interrogator questions two entities, $X$ and $Y$, and in the end must determine which one is a machine and which one is a human. Turing hypothesised that with the advent in designs of memory architectures, the interrogator would not have more than 70% chance to correctly identify the machine. John Searle proposed the *Chinese Room*, where the hand gestures of a Chinese language speaker could be interpreted as digital signals. If a computer could simulate those signals, thereby the hand gestures, that does not really conclude that the computer in itself is intelligent[196].

Thereafter, considerable progress had been made in the field of artificial intelligence. However, applications could not be implemented due to the lack of computing power. The processing and memory requirements of these tasks simply could not be fulfilled in those times. That changed in the 1990s when there was a leap in development of processor and memory architectures. There was an explosive improvement in computational power. More data became available and accessible to researchers. That resulted in breakthroughs in natural language processing, robotics, computer vision, and machine learning besides others. A number of early architectures and algorithms were developed throughout the 1950s and 60s. A mathematical model was published for building a neural network[197]. Later, the concept of back-propagation was floated that embodies today's deep learning driven AI systems[198]. Technology companies like Microsoft, IBM, and Apple started designing intelligent systems capable of handling millions of data and take credible decisions. One of the earliest examples is DeepThought, the computer program that (in)famously beat the then reigning world champion Kasparov in chess[199]. AI has been extensively used in the aviation industry. Craft flying into outer space have benefitted from AI as well. More recently, extensive research is being done in automated vehicles (self-driving cars). With the advent of huge amounts of processing power, image processing and recognition has taken a huge leap. Firms such as Google, Adobe, and a growing number of start-ups are harnessing the power of AI. Companies like NVIDIA are using their processing potential to generate language models (BERT, GPT-3) for accurate text classification[200]. In medical fields, AI is becoming instrumental in classification tasks and imaging[201].

## 4.2 Machine Learning

Machine learning is one of the categories of algorithms of artificial intelligence which tries to mimic human intelligence. One of the machine learning techniques is called deep learning. Machine learning, broadly speaking, constitutes of feeding data to a computer and learning from it, *i. e.* discovering heuristics, to get better at performing a task progressively. This technique renders the need of writing code redundant (or useless), but simultaneously obtains the desired outcome from the program, if not better (or an improved output). Deep learning takes the same data, assigns it weights, feeds it to a neural network architecture, where the data goes through multiple hidden layers, and creates relationships between the layers before producing the output with the best results[202–204].

Using pattern recognition and statistical methods, computers can perform specific tasks without being explicitly programmed to do so - that was the early theory of machine learning. Using AI, with all the bells and whistles, it is now possible to make computers learn from data, and with those heuristics it is also possible for machines to independently adapt to new, unseen data. The algorithms or techniques are inherently iterative, which enables the application to learn from previous computations to produce accurate and reliable results. This branch of computer science covers computers learning like humans by using algorithms to parse data and improving their learning over time in an autonomous fashion, and predicting or determining an outcome in the real world. It is the study of the way by which computer systems can improve with experience like humans do. The code gets shorter, the maintainability increases, and the program generally functions better.

The machine learning discipline harnesses statistics to find patterns in massive amounts of data by building a mathematical model based on the data. The data can be anything from words, numbers, images, and so on. Search engines, voice assistants, recommendation systems are based on machine learning. For example, when an image is captured by a smartphone camera, the way by which Google Photos tags the image is based on machine learning and is called image recognition. It trains a classifier to identify faces in the images and asks the user for a name to identify one image. If it is a person whose photograph was taken, the application searches through the faces and automatically tags the new photograph with the name of the person, if a previous photo exists in the gallery. Similarly, if the photograph is of a pineapple or ratatouille, it is identified as food.

Machine learning typically builds a mathematical model based on input data, also known as `training data`, which contains a set of data points, called `training instances or samples`, based upon which the model learns and performs a task of either determination or prediction. The task may fall into the category of *classification, regression, clustering*, etc. There is an `objective` or a `scoring` function which is used to evaluate the relationship between the data points defined in the model[203]. The parameters of the model are `fit` on the training data and observed if the desired outcome is achieved with the help of further parameter tuning and feature selection procedures. To determine the accuracy of the fitted model, it is then evaluated against test data, which contains exclusive data points unseen by the model, similar to the

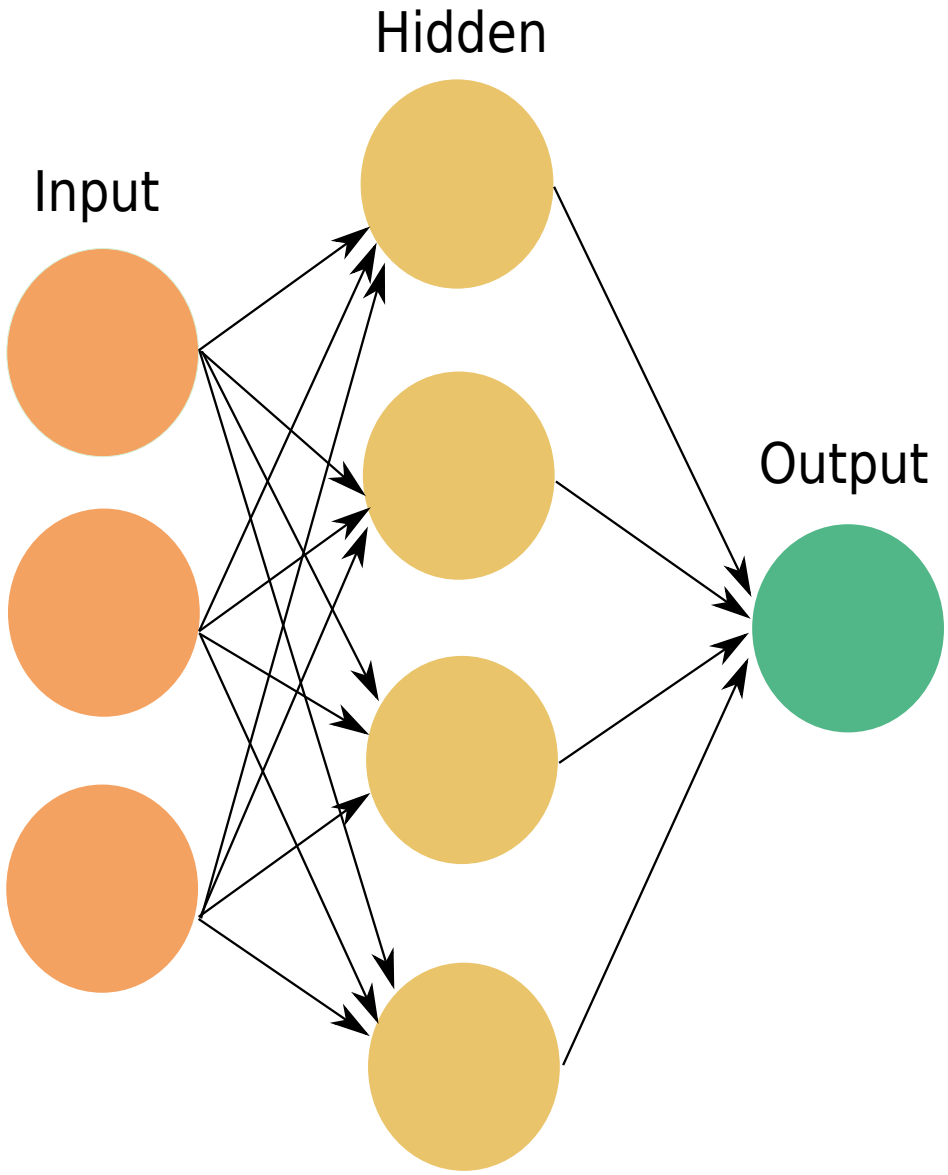



**Figure 4.1:** A simple neural network. In a feed-forward network the output from the previous layer is fed as input to the next layer, whereas backpropagation works backwards from the output layer to calculate the gradient.

ones in the training data, but not inclusive. The model is optimised by performing repeated searches or fits, for instance, to achieve the aim of generalisation so that it successfully interprets data beyond the training set. For a task `T`, if a machine performs at performance `P` and improves, learning from its experience `E`, then it can be remarked that this process belongs to the realm of machine learning[204].

Deep learning has several advantages over classical learning algorithms[205;206]. It essentially employs multi-layer neural networks mimicking biological networks for a wide array of tasks; although supervised problems can be solved by deep learning, unsupervised learning problems is the area where it excels as it intuitively gathers the required features and patterns to solve a task[205]. The features extracted during training are transformed into a distributed representational space, creating more combinations than was provided or learned - $n$ binary features can give way to $2^n$ different combinations, for instance. Layers of representation provides a deep learning

algorithm with exponential increase in depth, useful to extract even more meaningful patterns in data to be able to generalize. In a neural network (Figure 4.1), the hidden layers learn to represent the input layers paving the way to easier prediction of the output layers[206]. Multi-layer networks comprise layers, where one layer receives as input the output of the previous layer. Backpropagation enables an algorithm to compute the gradient of a neural layer or module in the reverse direction by starting with the output of a layer and moving backwards to the input. The gradients can be adjusted with respect to the weights of each layer[198]. The preferred architecture for deep learning tasks are the feed-forward neural nets, where algorithms map a fixed-size input to a fixed-size output by passing a weight sum of the inputs from the previous layer through a non-linear function. The most used function is rectified linear unit (ReLU) and is given by $f(z) = max(z, 0)$[207]. Deep feed-forward neural networks have found applications in speech recognition, natural language processing, image classification, and object recognition among several others[200;208;209]. Examples of deep learning architecture include convolutional neural networks (CNN) (explored below), recurrent neural networks (to process sequential data), and long short-term memory networks (LSTM) (deploying hidden units for long-term memory storage between layers)[206;210].

### 4.2.1 Components of machine learning

Machine learning contains various components that enable a model to successfully perform a task. The tasks may include object recognition, which is a form of classification, transcription of information into discrete textual format[211], anomaly detection, where a machine detects unusual patterns in data[212], and machine translation[200;208;213], among others. To perform the tasks, a machine also needs to consider several elements, some of which will be described below.

#### Training

Just like a toddler starts to learn and remember real world objects, algorithms follow a similar strategy to find out connections between various features and classes. Algorithms are exposed to datasets where they can learn the attributes of the data points and look for patterns. To start with the training process, the data points, numerical or categorical, are converted into smaller constituent units, which enables the learning algorithm to assess the underlying properties of the samples and features, and extract meaningful patterns. The data points are normally arranged as either *scalars* that are numbers denoting a certain property, such as frequency, or *vectors* that are one-dimensional arrays of scalars, or *matrices* denoting two-dimensional representations of the data points, or *tensors* that store multi-dimensional representations of data points. To convert them into more processable parts, the data points need to be decomposed. Among various options, *eigendecomposition* and *singular value decomposition* could be chosen for the purpose[214]. Eigendecomposition generally refers to solving a matrix for its eigenvalues and eigenvectors. It helps converting a matrix into a lower feature space, where the matrix can be scaled in any direction/dimension. In theory, if a non-zero vector $v$, when multiplied with a square matrix $A$, results in the same vector $v$, rescaled, given by a

coefficient $\lambda$, then $A$ has an eigenvalue $\lambda$ and an eigenvector $v$. It can be given as:

$$Av = \lambda v.$$

The eigendecomposition of $A$ can be defined as:

$$A = Vdiag(\lambda)V^{-1},$$

where $V$ is a matrix of eigenvectors $= [v_1, ..., v_n]$ and $\lambda$ is a vector of eigenvalues $= [\lambda_1, ..., \lambda_n]^T$. For a real symmetric matrix, an eigendecomposition would look like:

$$A = Q\Lambda Q^T,$$

where $A$ is the matrix, $Q$ is the orthogonal matrix if eigenvectors of $A$, and $\Lambda$ comprises of eigenvalues of $A$ as a diagonal matrix[214].
The eigendecomposition is not possible for all real matrices, however, singular value decomposition, or *SVD*, is. A matrix can be converted into singular values and singular vectors, regardless of whether it is square or not, unlike in eigendecomposition. Nevertheless, it is similar to the former method and can be denoted as:

$$A = UDV^T,$$

where $U$ contains left-singular vectors and $V$ has right-singular vectors as columns, both being orthogonal matrices. $D$ is a diagonal matrix whose diagonal elements are singular values. $U$ can be described as the eigenvectors of $AA_T$, while $V$ contains the eigenvectors of $A_TA$. The eigenvalues of $AA_T$ and $A_TA$ are in turn the squares of the non-zero singular values of $A$ populating $D$[214].

### Classification

For a set of data points, the system learns the patterns from the features and attempts to group instances from an unseen dataset into those targets. The machine learning algorithm is trained on data with labels (classes) and then tries to classify the new instances. Classification can be defined as a model categorising an input $x$ into a group out of $c$ different groups. The output label $y$ is deduced from $f(x)$, where $f$ is the learning function given by:

$$f : \mathbb{R}^n \to \{1, ..., c\}.$$

Apart from categorising an input into a class, classification can also included tasks such as a probability distribution over classes[214]. Spam filtering is an example, so is object recognition and, by extension, image recognition[215]. A different type of classification task arises when several learning functions need to be defined to extract patterns about an input whose feature subset is absent. That means, for every input $x$, a function learns from a subset of variables attached to $x$. To achieve classification accuracy, a probability distribution over all the qualifying features is considered before pruning the absent features. $2^n$ learning functions can be defined for $n$ inputs and only one of them is necessary to correctly produce the probability distribution and concurrently, classify the inputs[216].

### Regression

Instead of classifying, the algorithm attempts to predict a numerical value for the target[217]. To predict the price of a new product, for example, it learns from the features of the same family of products presently in circulation. Although the procedure is similar to classification, its output is different. The learning function can be defined as:

$$f : \mathbb{R}^n \rightarrow \mathbb{R}.$$

In linear regression, for instance, the function $f$ leverages the input vector $x \epsilon \mathbb{R}^n$ to predict the output $y \epsilon \mathbb{R}$, which is a linear function of the input. The prediction is given by:

$$\hat{y} = wx,$$

where $w$ is a vector acting as a coefficient to every input feature $x$ affecting the importance of the feature $x$ in the prediction: the larger the value of $w$, the higher the importance of $x$. $\hat{y}$ stores the predicted result. There can be an additional parameter, called the bias parameter, added to the regression function. This parameter $b$ is used for the transformation of the function and can be interpreted as the value the function is biased to, when there is no input (or when the features are irrelevant).

$$\hat{y} = wx + b$$

Increasing the feature space, the above equation could also be rewritten as:

$$\hat{y} = \sum_{i=1}^{n} (w_i x^i) + b$$

with $n$ being the total number of elements in the feature space[214].

### Feature engineering

While learning, the algorithm must extract patterns from the features of the data. Efficiently choosing those patterns is called feature engineering. This is important to identify patterns to distinguish between classes. It can be designed manually, but automatic feature extraction methods are preferred, the features selected generalize well[206].

### Cross-validation

After a model has been trained, the accuracy of the model can be determined using a validation set. Cross-validation refers to splitting the training set $x$-fold, retraining the algorithm on randomly chosen $x_n - 1$ folds $x$ times, and using the left out fold as the validation set to test the accuracy. With higher values of $x$ it is possible to observe

more variance in the data, subsequently reducing the bias of generalisation. However, if the number of samples is less, then a high $x$ will negatively impact the estimate of the confidence intervals, as there would not be simply enough variation in the data[218].

### Hyperparameters

These are different parameters that are tuned and optimised to extract the best possible performance of a model. Hyperparameters can be set up to search for the optimum values in relation to each other and this step can be performed during cross-validation by estimating the generalization error and updating the hyperparameters[219].

### Optimisation

A machine learning model seeks to perform a task on unseen inputs in the test set by learning from the training set. The application of the learned knowledge is known as generalization[220]. When a model is trained, it tries to predict the inputs in the training set. A training error is recorded for every prediction to evaluate the prediction performance. Similarly, a generalization error is also calculated on the model's performance on the test set. The goal is to reduce both the errors to have a robust model. If the training and test sets are generated from a random probability distribution, the expected training error will be equal to the generalization error owing to the common origin of the datasets. However, in machine learning, the model is trained on the training set first, its hyperparameters are tuned, before it is evaluated on the test set. This leads to the generalization error mostly being greater than the training error, since the sampling processes of the sets differ. Minimising the training error and reducing the difference between training and generalization errors determine the effectiveness of a machine learning model.
To achieve a predicted value, which cannot be further optimised, of a feature or an input, estimators are necessary[214]. A point estimator of data points $\{x_1, ..., x_n\}$ can be defined as:

$$\hat{\theta}_n = g(x_1, ..., x_n),$$

where $\hat{\theta}$ is the point estimate of a parameter, whose optimum value is represented by $\theta$. The objective is to return a value of $\hat{\theta}$ that is close to $\theta$ to reduce the error (training or generalization).
In the absence of an input, a function may be *biased* to a certain value and it is defined as:

$$b(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) \; - \; \theta.$$

The closer the expectation over the data, the less biased is the estimator. If $b(\hat{\theta}_n) = 0$, then the estimator is said to be unbiased.
The variance of an estimator depicts how much the estimated values vary from the original input data points. Variance is normally intended to be low to have a better

prediction. It is normally denoted as $\sigma$ and is calculated as:

$$\sigma^2 = (\frac{1}{n}\sum_{i=1} nx_i).$$

To estimate the generalization error, the *standard error of mean* is calculated and is given by $\frac{\sigma}{\sqrt{n}}$. The standard error is utilised to calculate the probability of the true expected values in a given interval, since mean of the data points is a normal distribution. The cross-validation techniques can identify the optimal bias or variance to use to minimise the generalization error. The *mean squared error* is also used to compute the deviation of the estimated values from the input taking both bias and variance into consideration. It is defined as:

$$mse = b(\hat{\theta}_n)^2 + \sigma(\hat{\theta}_n).$$

To obtain well-performing estimators, the *maximum likelihood estimator* principle is applied, which returns functions that can be optimum for various models. The maximum likelihood estimate of a parameter approaches its true value as the number of training elements increase. To find a maximum likelihood estimate, an arg max function can be defined as:

$$\theta_{model} = argmax_\theta \sum_{i=1} n \log p_{model}(x_i; \theta),$$

where $p_{model}$ returns the probability of the estimate $\theta$ being true to the data sample $x_i$, $i\epsilon(1, ..., n)$ and the points where the probability is maximum are returned. This is useful to achieve a closely related predicted distribution as the original input data. Furthermore, maximum likelihood estimator has the lowest *mse* among other estimators, hence it is convenient for use in machine learning tasks[221;222].

### Regularization

It is another technique employed to reduce a model's generalization error[223]. It does not, however, affect the training error of the model. This is applied as a penalty to the cost function, when there are more than one functions to choose from for the learning algorithm. Not every function will be appropriate for the prediction task at hand, since they are dependent on the representation space where the input variables reside and the output dimensional space. For example, a penalty, such as weight decay, can be added to a linear regression problem to regularize a model over the function $f(x; \theta)$, which can modulate whether the model overfits or underfits. To have control over better fitness of a model over the training data, a function can be defined as:

$$J(w) = mse_{train} + \lambda w^T w,$$

where $w$ denotes the weight and $\lambda$ is a constant tuning the strength of weights; the weights are taken as they are if it is 0, but become smaller if it is set to a higher value. The regularizer in this case will be $\Omega(w) = w^T w$ and can be used to train a regression model in polynomial space, effectively tuning the ability of the model to

choose from an array of different functions to achieve the best possible solution to the task. Regularization attaches preferences to different solutions by means of a penalty (to be minimised) and the solution that is most appropriate is chosen. Similarly, various regularization approaches exist from which an approach is selected that is well suited to the task at hand, instead of one single all conforming solution[214;223].

### 4.2.2 Types of machine learning algorithms

Machine learning can be categorised into two broad groups, depending upon what kind of experience $E$ an algorithm is going to have.

#### Supervised Learning

The mathematical model is built on data comprising both inputs and outputs, i.e. the training set is `labelled`. The backbone of supervised learning methods comprises the concept of *showing* the machine what to learn. The computer system receives a dataset full of `training instances`, which contains `features`, which are characteristics defining the instances, and a `label` or `target`, which is the expected output for a particular training instance. The program learns the features and is able to classify the data points based on them. The output is a vector of scores, which indicates affinity towards a particular class. An objective function calculates the error between the predicted and the actual scores and based upon that, some weights can be adjusted to minimise the error[206]. Stochastic gradient descent (SGD) is sometimes employed to adjust the weights, where the average gradient is calculated for a set of data points on the outputs and errors, following which the weights are accordingly adjusted[224]. For example, the Iris dataset[225] is one prime example on which a machine learning algorithm can be trained. The dataset contains measurements of 150 different plants. There are four features: *sepal length, sepal width, petal length and petal width.* There are three flowering species recorded in it to which the plants correspond, *i. e.* the labels. An algorithm can be trained to learn from the measurements and can be used to identify other plants having similar measurements. A supervised learning algorithm could be imagined as: for a set labeled training data $(x_i, y_i)$, $x_i \epsilon X$ and $y_i \epsilon Y$, where $X$ and $Y$ are input training examples and labels (outputs), respectively, the algorithm attempts to learn the function:

$$f : X \to Y. \tag{4.1}$$

The input vector $x$ normally is associated with a value $y$ and the learning procedure attempts to learn $y$ from $x$. It estimates the probability of $y$; $p(y)$ being associated to $x$ through the joint distribution of $p(y|x)$. Actually, maximum likelihood estimate of the distribution for a parameter $\theta$ can be calculated to find the probability of $y$ being close to the true value of $x$. For instance, in a binary classification problem where the classes are labelled 0 and 1, an approach with logistic regression will result in:

$$p(y = 1|x; \theta) = \sigma(\theta^T x),$$

where the logistic sigmoid function decomposes the linear function to return its output

in the interval $(0, 1)$, which is the probability of how close a value of $y$ is to a certain class[214]. Typical supervised learning tasks include classification, regression problems, support vector machines, decision trees among others.

#### Unsupervised Learning

The training set is *unlabelled*. The program learns from the characteristics of the data points and infers useful properties of the dataset. It attempts to extract properties from a probability distribution $p(x)$ of an input vector $x$, and there exists no $y$, unlike in supervised learning tasks. Unsupervised learning tasks try to understand the data only from $x$, since the `target` values are missing from the training set. This approach is generally related to density estimation, denoising data in a distribution, or clustering similar inputs into groups. The model collects representative features from $x$ which preserves enough information about it, such that it is easier to identify $x$ later for clustering, etc. The information about $x$ can be decomposed into i) tiny, low-dimensional representations, ii) sparse representations, where the information is stretched along the feature space onto multiple axes making it high-dimensional, and iii) independent representations, where the information is rearranged such that the dimensions are statistically independent[226;227].

### 4.2.3 A few machine learning algorithms

There are plenty of machine learning algorithms to choose from. They do not adhere to a one-size-fits-all principle, but they can be tailored to different tasks according to their capabilities. Some algorithms will be discussed here that are relevant to this thesis.

#### Support Vector Machines

Support vector machines (SVMs) attempt to learn decision rules from labeled input data to classify novel data and are considered to be one of the best performing *out-of-the-box* classifiers[228]. An SVM splits each sample to an $n$-dimensional vector and maps that onto multiple $n-1$-dimensional hyperplanes, following which SVMs choose the hyperplane which shows the largest variance or separability between the classes. SVMs are primarily applied to binary classification tasks and have been widely used for pattern recognition tasks, along with image and text classification problems. Their usage have branched to biological fields as well, especially computational biology. An SVM predicts class identities based upon the output of a linear function $w^T x + b$. Since the classes are numerically labelled, a positive output of the function results in the prediction of class 1, while a negative output infers class 0. The transformed sample vectors are utilised to identify similarities between samples and those are known as support vectors. The method of transformation of each sample is known as a *kernel* and the prediction function of an SVM is given by:

$$f(x) = b + \sum_i \alpha_i k(x, x_i).$$

$k(x, x_i)$ is a kernel-based function that actually returns a dot product between all inputs $(x)$, while transforming the data points in a new feature space. $\alpha$ is a coefficient vector and is linear to $f(x)$. Any model non-linear to $f(x)$ can be optimised in order to efficiently converge; since only $\alpha$ is optimised, the decision function is linear, albeit in a different space than the optimisation algorithm. This whole operation is called *kernel trick*, which can also be explained as adding another dimension to the hyperplane to aid separability between the two classes of data.
A widely used kernel is called the Gaussian kernel or the radial basis function kernel and can be defined as:

$$k(u, v) = N(u - v; 0, \sigma^2 I).$$

$N(x; \mu, \Sigma)$ denotes the standard normal density of the kernel whose value decreases along lines emanating from $u$ in $v$ space. Essentially, this kernel creates a support vector of a sample $x$ with a label $y$ and is assigned to a class (either 0 or 1). If another sample $x_a$ is found close to $x$ based upon Euclidean distance, it is considered to be very similar, hence a large weight is assigned to the label $y$. In the end, the classification depends upon the weights attached to the training labels[214].
There are other kernels available to train SVMs, for example: linear kernel, polynomial kernel, sigmoid kernel, and so on. To tune a kernel, the hyperparameters normally available are the regularization parameter and gamma. The regularization parameter (also called C) can be tuned to a higher value, which would set the hyperplane margins to be narrow, potentially increasing classification accuracy, while a lower value results in broader margins, with a chance for a drop in classification accuracy. Gamma sets the reach of the influence of a training sample to the hyperplane. If gamma is set to low, samples farther from the hyperplane will be considered; however, if gamma set to a higher value, the samples closer to the hyperplane will be considered in the decision making process[214;228;229].

### Random Forests

Random forests are actually an extension of decision trees. Decision tree-based classifiers are normally fast. They traditionally employ a central axis projection technique by constructing a hyperplane dividing the lines that connect two data clusters and identify classes at each decision node. However, the decision tree approach is prone to overfitting. Random forest-based classifiers grow multiple decision trees by splitting the feature space into random feature sub-spaces and train the individual trees practically on different subsets of the training data. The final classifier combines the accuracy measures from all the trees, thereby avoiding overfitting, i.e., maintaining generalization accuracy[230;231].
Random forests are a part of ensemble training algorithms, where an algorithm does not depend upon one function, rather aggregates outputs of different functions and chooses the best solution from that output space. Random forests achieve exactly that purpose by growing separate decision trees and aggregating the outputs. Each tree acts as a classifier that casts a vote on a class for an input sample. A tree is grown through random sampling with replacement and without pruning on the input samples. At each node of the tree a best split is selected based upon a constant that is randomly selected from the number of input variables. Furthermore, this constant $m$ is used to optimise

the correlation between two trees and strength of each tree in the forest. An increase in correlation results in decreased classification accuracy, while increased strength leads to increased accuracy. Random forests can work with numerous variables without the need to delete any of those and can return an estimate of variable importance. Concurrently, they can work with missing data. They can also compute proximities for two pairs of cases useful in clustering[231].

Random forests can generate unbiased generalization error estimates by out-of-bag (oob) sampling, where any given tree is grown with two-thirds of the bootstrap sample. The unused cases are classified by the trees already grown. With $c$ being the dominant class predicted for a oob case (from one-third of the sample), the ratio of the number of times that $c$ is not correctly classified averaged over all classes is known as the oob error estimate. This is used to generate an unbiased generalization error estimate and also for feature importance estimates. Feature importance for variable $m$ is calculated as follows: The votes for the correct class for the oob cases are calculated, followed by those from a random permutation of values in $m$. The mean of the difference in those two votes over all trees results in the importance estimate for $m$. The estimate is further divided by its standard error to obtain a z-score, which is normalized and a significance score is attached to it, resulting in the feature importance. The *gini* impurity index offers the probability of mis-classification of a particular sample by randomly choosing a label from a node. The *entropy* index calculates the information gain based on a particular node. If all of the samples in a particular node belongs to the same class, the entropy would be zero. Gini impurity is calculated by:

$$G = 1 \ - \ \Sigma_{j=1}^{c} p_j^2,$$

where $p_j$ is the proportion of samples belonging to class $c$ for a particular node.
Entropy is calculated by:

$$H = 1 \ - \ \Sigma_{j=1}^{c} p_j log_2(p_j),$$

where $p_j$ is the proportion of samples belonging to class $c$ for a particular node and $p \neq 0$[230–232].

If there are missing values in the input set, random forest classifiers can either fill them up by the median of all values in class $c$, if the variable is not categorical, or by the most frequent non-null value in class $c$, if categorical. In a different strategy, for a missing continuous value $x(m, n)$, the filling is done by calculating the proximities between the $n^{th}$ case and the non-missing value case and averaging the values of the $m^{th}$ variable, if non-categorical, otherwise, the most frequent non-null value weighted by proximity is used. A forest is grown and the process is reiterated until it is optimised[231].

The balance between prediction errors is achieved by random forests through allowing lower error rates for larger classes than for smaller classes. Random forests can be utilised to assign weights to the classes; the error rate of a particular class can be reduced by assigning it a comparatively higher weight[231].

The discriminant function for random forests proposed by Ho[230] is given as:

$$g_c(x) = \frac{1}{t} \sum_{j=1}^{t} \hat{P}(c|v_j(x)),$$

where $g_c(x)$ is the decision rule for a point $x$ assigned to class $c$ for which $g_c(x)$ is the maximum. $\hat{P}(c|v_j(x))$ is the estimate of the $P(c|v_j(x))$ and is given by:

$$P(c|v_j(x)) = \frac{P(c, v_j(x))}{\sum_{l=1}^{n} P(c_l, v_j(x))}.$$

$v_j(x)$ is the terminal node for a point $x$ given by a tree $T_j$, where $j = (1, 2, ...t)$. The probability of point $x$ belonging to a class $c$, where $c = (1, 2, ...n)$, is calculated as the ratio between $x$ belonging to $c$ and all points belonging to $v_j(x)$ and is given by $P(c|v_j(x))$[230].

### Principal component analysis

PCA is a statistical technique to compress data by reducing the dimensionality of large datasets. It is used to create new non-correlated variables that maximise variance and simultaneously to minimise information loss. PCA can be adapted to be used on numerical data without any distributional assumptions about the data. The representation learned by PCA uncouples all connected elements, thereby making them statistically independent. Additionally, the representation learned has a lower dimensionality than the original input. It operates by creating a multi-dimensional matrix ($m$ x $n$, where $m$ stands for number of samples and $n$ for variables for each sample) with $n$ $m$-dimensional vectors. The motivation behind the matrix is to look for a linear combination of the matrix columns with maximum variance, which is obtaining a vector. A covariance matrix is generated using that vector combined with the largest eigenvalue associated with the vector. The linear combinations derived from the covariance matrix are termed as the principal components[233]. The input data $x$ is projected onto a representation $z$, such that the direction of the greatest variance is aligned with the new axes. This transformation is orthogonal and linear. PCA essentially reduces dimensionality of data while preserving as much original information as possible by reconstructing the data in a lower dimension. For every point $x_i \epsilon \mathbb{R}^{\times}$, a vector $c$ is searched for by PCA such that $f(x) = c$, *i. e.* the point is encoded. To deconstruct it, a decoding function is generated such that $x \approx g(f(x))$. For the multi-dimensional matrix $X$ with $(mxn)$, an unbiased sample covariance matrix can be defined as below, after centering the data by subtracting the mean from all the data points:

$$Var[x] = \frac{1}{m-1} X^T X.$$

$z = W^T x$ is a representation that can generated by PCA following a linear transformation, where $Var[z]$ is the diagonal. The eigenvectors $X^T X$ constitute the principal components of $X$. In the end, PCA attempts to organise the principal axes with regard to variance of the data in a different space (with reduced dimensionality) and separate the data so that the points are statistically independent[214;233].

### $k$-means clustering

$k$-means clustering creates $k$ clusters and arranges the different observations into those disjoint clusters. Every cluster contains a cluster centroid, *i. e.* the mean of the samples in the cluster, and the algorithm attempts to tie each observation to a cluster in a way that its distance is at minimum to the centroid. To put simply, a one-hot vector $h$ of $k$ dimensions is created to represent an input $x$ in a sparse representation. $h_i = 1$, if $x \, \epsilon$ cluster $i$, and all the other entries in $h$ will be rendered 0. This representation ensures all samples to be in the same cluster if they are similar and becomes computationally inexpensive. $k$-means algorithm chooses the nearest centroid for each input from a group of $k$ different centroids $(\mu_1, ..., \mu_k)$ and then updates the mean of every cluster $i$ and assigns that to each centroid $\mu_i$. Ultimately, it aims to cluster points into $k$ groups of equal variance. Being an unsupervised approach, $k$-means clustering deduces some similar properties about the input data to sort them into various clusters. Doing so, it might lose some relevant information and sort the data into wrong clusters. Besides that, it might also find some valid relationship between the features, which might, however, end up in formation of unexpected clusters[214;234;235].

### Convolutional Neural Networks

Convolutional neural networks (CNN) receive data as input in multiple arrays and use convolution instead of matrix multiplication in one of the constructed underlying layers[206;214]. These networks hark back to neuroscience, wherein the layers function akin to simple cells and complex cells. Like any other deep learning architecture, CNNs are multi-layered and are built on the principles of local connections, shared weights, and pooling[205;206]. The input can be represented either in 1D (for signals and sequences), 2D (for images), or 3D (for video). Convolutional layers and pooling layers comprise the physical structure of CNNs, each layer being subdivided into units belonging to various feature maps. The units accept a weighted sum generated from the outputs of the previous layer passed through a non-linear activation function (like ReLU). Each feature map has its own filter bank. The CNN design preserves the invariant nature of signals by implementing shared weights to different units in different feature maps carrying the same signals or patterns. The filtering operation is called convolution[206]. In its basic form, it can be defined as:

$$s(t) = (x * w)(t).$$

For a 1D input, $s(t)$ is the feature map for the input $x$ with $w$ being the weighted average, also known as the kernel. The convolution operation is given by the asterix. The input and the kernel are generally multidimensional arrays (tensors), where the kernel defines parameters to be used by the algorithm. The input is usually larger than the kernel, invoking sparse interactions between the layers by detecting small features among the layers, thereby reducing memory usage. For a 2D input (such as images),

the convolution can be defined by:

$$\begin{aligned} s(i,j) &= (K * I)(i,j) \\ &= \sum_m \sum_n I(i-m, j-n) K(m,n), \end{aligned}$$

where $I$ denotes the 2D input, $K$ denotes the kernel. $m$ and $n$ indicate the finiteness of the data, although infinite iterations can be achieved, distributing the representation of the feature space[214;236;237].

The convolutional layers detect local conjunctions of features from the previous layer, whereas pooling layers merge the features semantically[206]. The feature representation is shrunk to a lower dimension by the pooling layers by rearranging the unit patches based upon their maxima, albeit through small shifts and distortions. A CNN is based upon the compositional hierarchy of natural systems, where lower-level features combine to form higher-level features, which can found in images, texts and sound data. A pooling layer ensures that these hierarchies remain tractable by forcing representations to be invariant, such the outputs of pooling layers do not significantly change. Invariance to translation is the result of pooling and features can learn to identify transformations of convolutional layers which can be treated as invariant. It is useful to determine a feature in its expected location. Furthermore, pooling reduces the dimension of the representation, improving efficiency by enabling the following layer to process fewer inputs, also preserving the integrity of the data. There are may be multiple convolution and pooling layers, capturing the essence of a multi-layered architecture. Backpropagation is used to compute the gradients through both the convolutional and pooling layers[206;214].

CNNs have been deployed in various use cases, namely face detection, text recognition, traffic light recognition, and medical image detection[215;238–240]. Recent advances in technology have seen multiple ReLU layers being deployed, new regularization techniques such as dropout being used, GPUs being upgraded and put to more efficient use, essentially enhancing performance of CNNs to have billions of weights and unit connections[206;241].

### 4.2.4 Performance metrics

After training the model on the data, it is imperative to evaluate the performance of the model. Measurement of training error and generalization error have already been presented above. Another performance evaluation method would to calculate the mean squared error of the model on the test set. It can be defined as[214]:

$$mse\_test = \frac{1}{n} \sum_i (\hat{y}\_test - y\_test)_i^2,$$

where $y\_test$ is a vector with the actual test values (labels for a supervised task) and $\hat{y}\_test$ contains the predictions of the model. For a 100% accurate classification, $\hat{y}\_test$ will be equal to $y\_test$, rendering the error to zero. The equation below shows that on

increase of the Euclidean distance between the actual values and predicted values, an increase in error is observed[214]:

$$mse\_test = \frac{1}{n}\|\hat{y}\_test - y\_test\|_2^2.$$

To optimise $mse\_test$, some weights $w$ need to be assigned and tuned that will eventually reduce the error. If the error on the training set ($mse\_train$) is minimised, that will automatically induce reduction in $mse\_test$. A suitable $w$ can be obtained if $mse\_train$ is solved for where its gradient is 0 to acquire a simple learning algorithm[214]:

$$\nabla_w mse\_train = 0$$
$$w = (X\_train^T X\_train)^{-1} X\_train^T y\_train,$$

where $X\_train$ is the training set and $y\_train$ contains the labels.

The prediction that is obtained as the outcome, can be measured with various performance metrics. A brief overview of performance metrics used for a classification problem is given below, since they are more relevant to the work that will be presented later. To accurately visualise the metrics, a binary classification task was designed. A random forest classifier was trained on 1000 randomly generated samples using `make_classification` class of `scikit-learn`[242] with two features. The two classes were labelled as A and B. The sample set was further split into a training set and a test set with the ratio 0.8:0.2, *i. e.* 800 of the samples constituted the training set and the rest the test set. Following that, the random forest classifier was trained and the required metrics were generated.

```
X, y = make_classification(n_samples=1000, n_features=2,
                           n_informative=2, n_redundant=0,
                           random_state=42, shuffle=False)

X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

rf = RandomForestClassifier(random_state=42)
rf.fit(X_train, y_train)

classification_report(y_test,rf.predict(X_test))
```

- **Confusion Matrix**: A confusion matrix is perhaps the most useful tool to have a general visual overview of the measure of accuracy of the model. Although it is not a metric in itself, it is easier to define the metrics that come afterwards. It can be applied to a multi-class classification problem, besides a binary classification task, and it displays values belonging to each of the predicted and true labels. For example, in a typical image classification problem of distinguishing an aeroplane from a bird, which is essentially a binary classification task, a confusion matrix can be used to visualise the various metrics. For two classes, $A$ and $B$, the confusion matrix for this task will look like in Table 4.1.

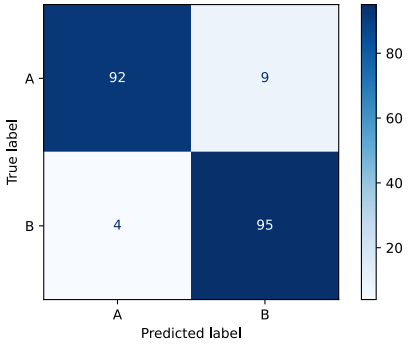



**Figure 4.2:** The confusion matrix

| | | Actual | |
|---|---|---|---|
| | | A | B |
| Predicted | A | TP | FP |
| | B | FN | TN |

**Table 4.1:** The confusion matrix

For the sake of clarification, the A label is assigned the value 1 and the B label the value 0. TP stands for *True Positive*, where both the actual and predicted labels are `True`.
FP stands for *False Positive*, where the predicted label is `True` or 1, but the actual label is `False` or 0.
TN stands for *True Negative*, where both the actual and predicted labels are `False` or 0.
FN stands for *False Negative*, where the actual label is `True`, but the predicted label is `False`. This is the case when an element of $A$ is misidentified as an object of $B$.
The objective of the model is to maximise the diagonal values (to bring them closer to 1) and reduce the non-diagonal values. Having 1 on both the diagonal elements is the (non-realistic) ideal case. This example can be extended to classification problems such as identification of a spam email, or diagnosis of a disease, where obtaining higher values on the non-diagonal axis can have detrimental effects, especially in the latter case. For the example classification problem with randomly generated samples, the confusion matrix is given in Fig 4.2.

- **Accuracy**: It gives the ratio of the number of correct predictions over all the predictions made by the model.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

The accuracy for the classification task is 94%.

- **Precision**: The ratio of the true positives among all the positive predictions is given by this measure. From the example, precision shows how many objects predicted to belong to $A$ (or $B$) actually belong to $A$ (or $B$) (how many predicted birds were actually birds)[243].

$$Precision = \frac{TP}{TP + FP}$$

- **Recall**: This measures gives the percentage of the `True` class that is predicted. In the example, recall is the measure of how many elements of $A$ were correctly

predicted from the predicted values also including elements of $B$ that were falsely classified as belonging to $A$[243]

$$Recall = \frac{TP}{TP + FN}.$$

Recall is also known as **sensitivity**.

- **Specificity**: It is the ratio of the true negatives over the wrongly predicted results and correctly predicted `False` classes. It will be the ratio of wrongly identified elements of $B$ over the total of correctly and wrongly classified elements of $B$ (with respect to $A$ and vice versa)[244]

$$Specificity = \frac{FP}{FP + TN}.$$

- **F value** It is the harmonic mean of precision and recall aiming to balance the two metrics[245]

$$F\,value = \frac{(1 + \beta^2).precision.recall}{precision + \beta^2.recall}.$$

where $\beta$ shows the relative importance of precision vs recall and is normally set to 1[246]. All the scores of the evaluation metrics for the example classification problem can be seen in Table 4.2.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| A | 0.96 | 0.91 | 0.93 | 101 |
| B | 0.91 | 0.96 | 0.94 | 99 |
| accuracy | | | 0.94 | 200 |
| macro avg | 0.94 | 0.94 | 0.93 | 200 |
| weighted avg | 0.94 | 0.94 | 0.93 | 200 |

**Table 4.2:** Summary of metrics of the example. The support column indicates the number of true values were there for each class for the upper rows, total test cases for the lower metrics.

- **ROC curve** and **AUC**: The receiver operating characteristic (ROC) curve for a binary classifier shows a curve of its true positive rate against its false positive error rate rate based upon several thresholds (a cut-off threshold being the probability beyond which the classifier makes a decision on a label, which can be altered as a parameter). The increase in accuracy of positive samples can be achieved at an increased cost of accuracy of negative samples and the relationship between the two elements is defined by ROC curve. The x-axis shows the negative sample error rate, while the y-axis shows the positive accuracy. The area under the curve (AUC) aims to aggregate the performance on all possible thresholds and gives an overview of the precision and recall metrics. The bigger the area, the better the performance of the classifier[247].
  For the binary classification task at hand, the ROC curve is given in Fig. 4.3.

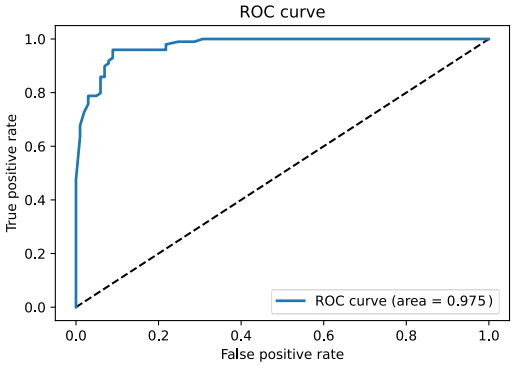



**Figure 4.3:** ROC curve. The ROC curve and the AUC. The coordinates for the best possible result is (0,100).

- **MCC**: The Matthews correlation coefficient was posited initially to compare chemical structures and has been applied in the machine learning context for binary classification tasks and can be extended to multi-class classification tasks as well[248;249]. This metric returns a value between $-1$ and $+1$, and does it only if the binary classifier correctly predicts majority of positive and negative data points. MCC is reliable in tasks where datasets are imbalanced. Moreover, in contrast to F value, it considers correctly classified negative classes and is invariant towards class swapping[249]. MCC is given by:

$$mcc = \frac{TP.TN - FP.FN}{\sqrt{(TP+FP).(TP+FN).(TN+FP).(TN+FN)}}$$

  The MCC for the example classification task is 0.871.

- **Cohen's kappa score**: The Cohen's kappa score provides a measure to test inter-rater reliability. It is implemented to evaluate something more than the simple accuracy of a classification problem, rather the level of agreement between two classifiers (or annotators) by chance. However, the kappa score can be misleading owing to its high sensitivity towards the distribution of marginal totals[250;251]. The metric can be defined as:

$$\kappa = \frac{(p_0 - p_e)}{1 - p_e},$$

  where $p_0$ is the empirical probability of agreement on a label and $p_e$ is the expected agreement. Good agreement is generally shown by scores above 0.8, whereas 0 or lower scores means random prediction[242].
  The Cohen's kappa score for the example classification task is 0.87.

- **Rand index**: For clustering problems, rand index measures the similarity between two classes. Its definition is similar to the accuracy metric for supervised classification tasks, but is also applicable in tasks where label information is not available, e.g. in $k$-means clustering. It returns scores in the range of [1,-1], where scores close to 1 signify possibility of identical clusters[252]. It can be defined as:

$$r = \frac{a+b}{C_2^{n_{samples}}}.$$

$C$ refers to the training (labelled) data and $C_2^{n_samples}$ is the total number of possible element pairs. Assuming $K$ to be the clustering, $a$ refers to the total number of pairs that belong to the same set in $C$ and to the same set in $K$, and $b$ stores the number of pairs that belong to different sets in both $C$ and $K$[242].

### 4.2.5 Challenges

To bring out the best performance of a model, the machine learning algorithms must overcome certain challenges, one of the challenges being *overfitting*. The model tends to show a bias towards the training data and does not generalise towards new data. This implies that the gap between the training error and generalization error is too large. Contrary to overfitting, *underfitting* models do not learn the important patterns from the entire range of the features and tend to make similarly false assumptions. When there is a large error returned on the training set, contrary to a desired small error value, the model is said to be underfitting. To achieve a better degree of control, a set of functions can be made available to the model, called the *hypothesis space*. The model can choose a function which is more appropriate for a given sample space; for example, selecting a polynomial function over linear function gives a model finer controls over its learning abilities. By modulating the hypothesis space, a model's capacity can be tuned to be appropriate for the complexity of the task. A model tends to overfit if its capacity is high and it is too influenced by the patterns of the training set to extract any discernible patterns from the test set. It tends to underfit if its capacity is low and, as a result, it cannot efficiently glean information from the training set. This can be rectified by altering the number of input features, for example. The representational capacity of a model determines a group of functions with varying parameters to choose the best function from, but often the learning algorithm chooses the function that significantly reduces the training error, which might not be the best function. Owing to the difference in training set sizes, training and generalization errors vary; generalization error stays the same or decreases with increase in training data. For a model with optimal capacity, it is possible to observe a large gap between training and generalization errors, which might be mitigated using a larger training set. *Curse of dimensionality* is another issue that arises where patterns found in higher dimensions, with more features, cannot be replicated in lower dimensions, with fewer features. The data points in higher dimensions are sparse[203;246;253;254].

Additionally, datasets may suffer from over-representation or under-representation of a certain class. This can happen when there are manifold more samples of a certain class than the other one in a binary classification problem, for example, and that affects the performance of the model. In this case, the dataset is called *imbalanced*. A model trained on this dataset detects the class that is over-represented and reaches a decision based upon that information, ignoring the other class. In image or text recognition problems, this issue has been observed to be prevalent. If images of a certain label $i$ populate the dataset over any other categories of images, a high number of the images, if not all, might be classified to belong to $i$, resulting in false positives. This abundance of elements from one class induces a bias to that class while classification. To evaluate a model trained on imbalanced datasets, classification accuracy is not the

most appropriate performance metric, since it will be dominated by the number of true positives (from the over-represented class) and will tend towards 1.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

As there are fewer objects from the under-represented class, they will possibly be identified as false positives (instead of true negatives) and won't contribute to the accuracy measure. ROC curve and area under the curve can be deployed to measure the performance of the models, as they represent the relationship between the true positives and the false positives. They will return a more comprehensive result than predictive accuracy rates as they do not consider entire output space of the model[246;247;255]. Similarly, precision and recall metrics can also enable the model to improve its learning from the dataset.

$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$

The trick here is to increase the recall of the learning algorithm without having an impact on the precision, which is difficult, since the number of false positives might increase while increasing the true positives of the under-represented class. The F value can combine the trade-offs of both precision and recall to produce the efficacy of the classifier[243;246].

As already mentioned, the imbalance in datasets can stem from over-representation or under-representation of a single class. It can also be attributed to sparsity in feature space or distribution of data within each class[256]. To solve this problem, various sampling approaches have been proposed. One of the approaches is over-sampling with replacement, where elements are added to the under-represented class (mostly copies). Under-sampling refers to deleting random elements from the over-represented class to match or get close to the under-represented class. They are great approaches to attack the imbalanced datasets issue, but under-sampling can result in removal of important information from the dataset, while over-sampling can lead to over-fitting. It is possible to selectively remove elements based upon a one-sided selection approach proposed by Kubat and Matwin[257]. Focused under-sampling refers to removing the elements of the over-represented class that do not occur in vicinity of the decision boundary of the two classes. In contrast, focused over-sampling adds copies of vicinal elements of the decision boundary of the under-represented class[246;256]. An over-sampling technique called SMOTE (for Synthetic Minority Oversampling Technique) for generating synthetic samples is applied to the feature space of the under-represented class to achieve larger and less specific decision regions for the learning algorithm, as opposed to over-sampling with replacement, where decision regions are smaller and specific. Different ensemble-based learning algorithms have also been observed to perform well with imbalanced datasets. Attaching cost penalties to algorithms can also be a solution, where they can be forced to focus on under-represented classes by penalising them for false positive errors. Finally, decomposition of the over-represented class into multiple classes based on comparatively more similar properties could help the learning algorithm extract

improved information[246]. Classification without any bias towards a particular class is the ultimate aim.

Another issue that can lead to imperfect classification in machine learning is data sparsity. Datasets can have missing values simply due to unavailability of data or it could be because of more complicated reasons, such as corrupt measurements. In machine learning, this problem can be alleviated by a technique called data imputation which employs statistical approaches to estimate a value from the values present and replace the missing values with the estimate. In deep learning, which relies on enormous amounts of data to perform a task efficiently, loss of data becomes crucial. To address that, different models have been proposed based on training generative deep models through an adversarial process called GANs. GANs are best deployed in image completion, where they are trained on real images to be able to construct a fake image by almost realistically reconstructing the missing pixel data distribution[258].

# III

# Exploring the transcriptome

# 5

# Splice variants in lncRNAs

A study conducted to better understand the lncRNA transcriptome, specifically splicing mechanisms, will be described in this chapter. The motivation behind this study was to comprehend the completeness of lncRNA annotation, given that it still lags behind annotation of protein-coding genes.

## 5.1 Presence of rare isoforms

It has already been observed that long non-coding RNAs are an integral part of the mammalian transcriptome, especially the human transcriptome[101;259]. They are also involved in a wide variety of regulatory mechanisms. Compared to protein coding genes, they are often expressed at low levels and are restricted to a narrow range of cell types or developmental stages, but they are evolutionary conserved, at least across evolutionarily closer relatives[2;116;160]. Several roles of lncRNAs include chromatin modification by acting as scaffolds or guides to ribonucleoprotein (RNP) complexes, RNA-mediated decay, and acting as decoys or sponges targeted by miRNAs for post-transcriptional gene regulation as already discussed in Chapter 3. Although the set of lncRNAs that is well understood with respect to biological function and molecular mechanisms is still limited, it is rapidly expanding through experimental and computational analyses, given a growing interest and widespread access to high throughput sequencing technology. Nevertheless, the coverage and precision of the lncRNA annotations lag behind the accurate maps of protein-coding genes. As already reported, lncRNAs genes are multi-exonic and produce multiple isoforms. There is evidence of alternative splicing and of one dominant isoform, which constitutes the bulk of the gene's expression[24;116]. Concomitantly, the diversity of their isoforms is still far from being recorded and catalogued in its entirety, and it remains to be seen what fraction of non-coding RNAs truly conveys biological function rather being just transcriptional noise.

The GENCODE project provides the most accurate transcript and gene annotation for the human genome[5;56]. It is a combination of manual and automated annotation techniques which endeavours to list gene features from HAVANA and Ensembl datasets[260]. Detailed surveys of expression patterns across many tissue and cell types evince intricate regulatory networks in which lncRNA genes are key players[76;121;261]. There is mounting evidence, however, that lncRNA isoforms may differ drastically in their biological function. For example, the lncRNA *ANRIL* has been known to suppress *CDKN2A* and *CDKN2B* in *cis* in prostate tissue. Contrastingly, its isoforms have been shown to regulate *TSC22D3* and *COL3A1* in *trans*[262]. *LINC00663* was also found to be differentially expressed in several cancer cell lines[263].

Historically, lncRNA gene models were often truncated due to their low expression values. Hon et al.[76] aimed to provide specific 5' end maps of lncRNAs. The situation is still more difficult at the 3' end, since long unspliced 3' end regions make it difficult to determine complete transcripts from Illumina data[264;265]. Furthermore, lncRNAs such as *ANRIL* exhibit complex patterns of alternative splicing.[262] Even in extremely well-studied protein-coding loci, rare isoforms keep being discovered[39]. Thus, the question of the extent of completion of the current maps of lncRNAs remains unresolved, both in terms of the number of expressed transcripts per gene and in terms of variability of their isoforms.

## 5.2 B-cell lymphoma

The idea behind this research work was to look for answers to two specific questions. Firstly, it was to be determined to what extent the transcript portfolio of a particular cell type has been mapped in its entirety. Secondly, in relation to the solution to the first question, the percentage of reported transcripts was noise will be explored. To address these issues, a very large set of transcriptome data from B cell lymphomas was investigated. The motivation behind this idea was that by virtue of aggregating hundreds of independently generated transcriptome datasets, it could be studied in detail, if the set of detectable splice junctions converged to a consensus.

To implement this idea, RNA-Seq samples from the Molecular Mechanisms in Malignant Lymphomas by Sequencing (MMML-Seq) project, specifically focused on B-cell lymphomas and part of the International Cancer Genome Consortium (ICGC), were collected[6]. B-cells are a type of white blood cells activated as a humoral immune response when they bind to an antigen, predominantly present in a lymphoid tissue, and play a crucial role in cancer checkpoint blockade therapies[266]. B-cell lymphomas is the group of cancer that is proliferate when lymphocytes up-regulate B-cells. Most B-cell lymphomas are classified as non-Hodgkin lymphomas, however some are Hodgkin lymphomas. The most common type of non-Hodgkin lymphoma is *diffuse large B-cell lymphoma* (DLBCL), constituting over 30% of all B-cell lymphoma cases, that is normally divided into germinal centre B-cell like (GCB) and activated B-cell like (ABC) subtypes based upon their biogenesis[267]. The other types include, but not restricted to, *follicular lymphoma*(FL), which is a lymphoid tissue neoplasm, also showing evidence of germinal centre B-cell differentiation[268], *high-grade B-cell lymphoma* that exhibits MYC and

BCL2 and/or BCL6 rearrangements[269]. and *Burkitt lymphoma* (BL), the latter being a more aggressive type of B-cell lymphoma, down-regulating MYC and translocating immunoglobin-MYC (IG-MYC) complex, facilitating ID3 gene inactivation[6].

### 5.2.1 Data processing

#### Annotated data

To analyse the annotation landscape of lncRNAs, detailed annotation data from the various publicly available annotation catalogues were considered. They included data from GENCODE releases 7 through 24[270], which were to be used as the benchmark for this analysis. Ensembl releases 60 and 83[271], NONCODE 2016[272], and the annotation catalogue curated by Cabili. et al.[121] were chosen purely to observe the transcriptomic topography of annotation provided by other researchers. The then available categories "antisense", "lincRNA", "processed_transcript", and "sense_intronic" were accepted as lncRNA biotypes (Fig. 5.1). The annotation data was used to define the location of the individual genes to count the total number of introns. An in-house pipeline[66] written in the `Java` programming language was used to aggregate the multiple transcriptomes and compute summary statistics of interest. Input datasets were provided in the standard Gene Transfer Format (GTF) files, which contains fields indexing various features of each gene, viz. its chromosome, genomic coordinates, strand specificity, as well as its constituent transcripts and their exon related information.

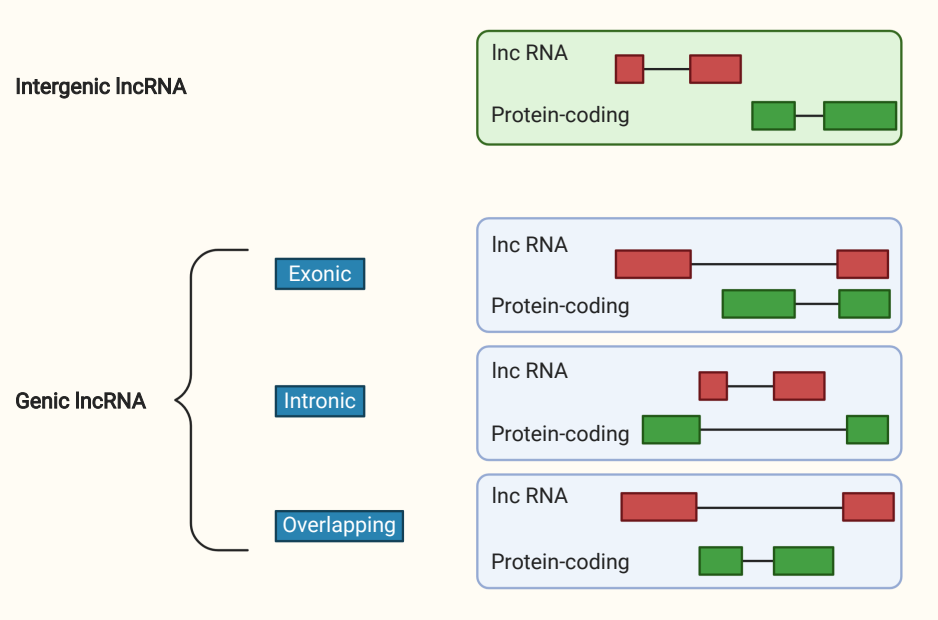



**Figure 5.1:** GENCODE v7 biotypes. The lncRNA biotypes as first defined in GENCODE v7[56]. The lncRNAs can be divided into intergenic and genic lncRNAs. Genic lncRNAs contain exonic lncRNAs, which overlap with an exon of a protein-coding gene; intronic lncRNAs, which reside in the intronic region of a protein-coding gene; overlapping lncRNAs, which overlap a protein-coding gene in sense.

Within a gene structure, transcripts typically overlap and share exons. Therefore, the set of unique exons was determined for every gene, which was subsequently used to

calculate the number of unique introns. For every gene, the average number of exons and introns were calculate from the set of unique exons and introns present - without considering how often they appear in distinct transcripts.

### Independent RNA-Seq data

From the MMML-Seq project data, 111 RNA-Seq samples were collected, that included samples from BL, FL, and DLBCL. The short-read sequencing data from the samples was mapped onto the human reference genome `hg19` using the splice-aware mapping tool segemehl v.0.1.7.[39;273] Based on enhanced suffix arrays (ESA), segemehl aligns sequence reads to a reference sequence. From the read provided, the tool scores the seeds and passes the best scoring seed to the next level, which a semi-global alignment procedure. It also entails a split-read mapping strategy that enables the user to have circular alignments or alignments in *trans* besides reads overlapping multiple reference sequences. The tool does not consider any existing annotation information before computing a splice junction, hence the association of a splice junction to a gene in mostly unambiguous. The mapping algorithm indirectly favours canonical splice junctions[33], since the alignment score for read fragments ending at these sites is higher.

```
segemehl.x -x <hg19>.idx -d <hg19>.fa

segemehl.x -S -i <hg19>.idx -d <hg19>.fa
-q <bcl_short_reads>.fa > <aligned_map>.sam

haarz.x split -m [<1,5,10>]
-f <aligned_map>.sngl.bed > <split_reads>.bed
```

The example commands above show a bare-bones approach to performing the mapping of the short-read sequences on the reference genome. The first command indexes the reference FASTA file which is the fed into the second command, triggering the mapping operation, enabling the split-read consideration capability the tool with the `-S` switch. The number of reads amounted to around 120 million per RNA-Seq sample, i.e., more than 10 billion reads. The length of each read was 101 nucleotides and around 90% of them could be mapped. The read supports of all genomic intervals spanning exon-exon boundaries, called splice junctions, were calculated. The genomic regions surrounding the junctions were branded as potential introns. To call the fragment an intron, each junction was required to have a minimum read support of one, five, or ten reads representing it, specified by the `-m` switch in the third command.

### Comparison

To better understand the distribution of splice junctions across the lincRNA genes, the mapped reads were further compared with available annotation data. It is to be noted that since the RNA-Seq data is not strand-specific, only non-overlapping, or intergenic, lncRNA (lincRNA) genes were considered for this analysis. In order to address changes in the annotated gene structure over time, GENCODE v.19 genes were chosen as

reference, as genes from this GENCODE version had the highest overlap with the lymphoma samples, as opposed to the other datasets. The intersection of GENCODE v.19 with the mapped loci in the lymphoma dataset that are supported by at least 10 reads resulted in 5,257 lincRNAs. As segemehl identifies split-reads independent of any annotation, all splice junctions located within the genomic coordinates taken from GENCODE were considered as potential introns belonging to that particular GENCODE annotated gene. This procedure was repeated over the complete range of the scope of the samples on all GENCODE releases. A summary of the overlap between lincRNAs of the lymphoma dataset and GENCODE annotations can be found in Table 5.1.

**Table 5.1:** Overlapping lncRNAs in lymphoma dataset and GENCODE. Overlap between lincRNAs expressed in the lymphoma dataset and different versions of the GENCODE annotation.

| | Genes | Transcripts | Exons | (avg) | Introns | (avg) |
|---|---|---|---|---|---|---|
| v7 | 3,296 | 4,563 | 12,584 | 2.76 | 8,394 | 1.84 |
| v19 | 5,257 | 7,487 | 18,774 | 2.51 | 12,010 | 1.60 |
| v24 | 4,961 | 7,318 | 18,685 | 2.55 | 12,202 | 1.67 |

## 5.3 More splice variants found

A comparison of published annotation data showed substantial differences in the average number of exons in a transcript, with some systematic trends over time. Ensembl version 83[274], which shared GENCODE version 24 as basis of gene annotation, reported slightly fewer introns for every lncRNA gene than earlier versions. Concomitantly, earlier Ensembl versions reported only a limited number of lncRNA genes. For instance, Ensembl v.60[275] included only 1,443 lncRNAs compared to 15,941 lncRNAs in GENCODE v.24. The mean introns present were rather close to the genomic data compiled in the much more complete GENCODE v.7 annotation. In contrast, the NONCODE database proved to be very inclusive (and still is[11]) and provided more than an order of magnitude more entries. Correspondingly, the GENCODE annotation also exhibited a moderate decrease in the number of exons for lncRNA genes over time, as can be seen in Table 5.2. This was thought to be a consequence of the fact that more recently included lncRNA could contain a larger fraction of single exon transcripts.

**Table 5.2:** Mean exons and introns across annotation catalogues. lncRNA genes catalogued by various annotation systems. The average number of exons and introns per transcript is given in the *(avg.)* column.

| | Genes | Transcripts | Exons | (avg.) | Introns | (avg.) |
|---|---|---|---|---|---|---|
| Ensembl 60 | 1,443 | 1,703 | 4,921 | 2.89 | 3,218 | 1.88 |
| Cabili 2011 | 8,263 | 14,353 | 33,045 | 2.30 | 18,607 | 1.30 |
| NONCODE 2016 | 16,0376 | 233,696 | 536,111 | 2.29 | 305,771 | 1.31 |
| GENCODE v7 | 9,580 | 14,984 | 42,060 | 2.81 | 28,998 | 1.94 |
| GENCODE v24 | 15,941 | 28,031 | 68,457 | 2.44 | 45,016 | 1.61 |

Subsequently, a systematic investigation of the influence of the dataset size on the complexity of inferred gene structures was undertaken to better comprehend the structure of rare isoforms in the human transcriptome. Analysis of the RNA-Seq data and its subsequent overlap with the GENCODE annotation yielded some interesting results. Figure 5.2 summarises the effect of increasing coverage on the estimated average number of introns per gene locus. Here, only the subset of expressed lincRNAs with at least one annotated intron was used to reduce the probability of erroneously or ambiguously mapped reads, which reduced the genes to 1,441. The data qualitatively reproduced the observation of the ENCODE project that there was a large difference in the average number of splice junctions between protein-coding loci and ncRNAs[56]. Although restricted to a quite narrowly defined cancer type, the average number of introns in lincRNAs was observed to be greater by about one than found in the GENCODE dataset, which uses a composite of a broad range of cell lines an tissues. An analysis of GENCODE v.19 revealed an average of 3.21 introns for these lincRNAs, whereas the analysis of the RNA-Seq data disclosed a mean of 4.29 introns for the same lincRNAs. Moreover, slightly more than 8% of the introns were noted to be novel, which can be positively attributed to the effect of the extreme sequencing depth of the combined lymphome data. The curves in the plot also show that the data saturated very slowly, requiring dozens or even more samples to reach the plateau value. The data showed that the detected splice junctions were unlikely to be noise since the curves saturate well for all three minimum read support thresholds [1,5,10] instead of showing linear growth, affirming their physical reality.

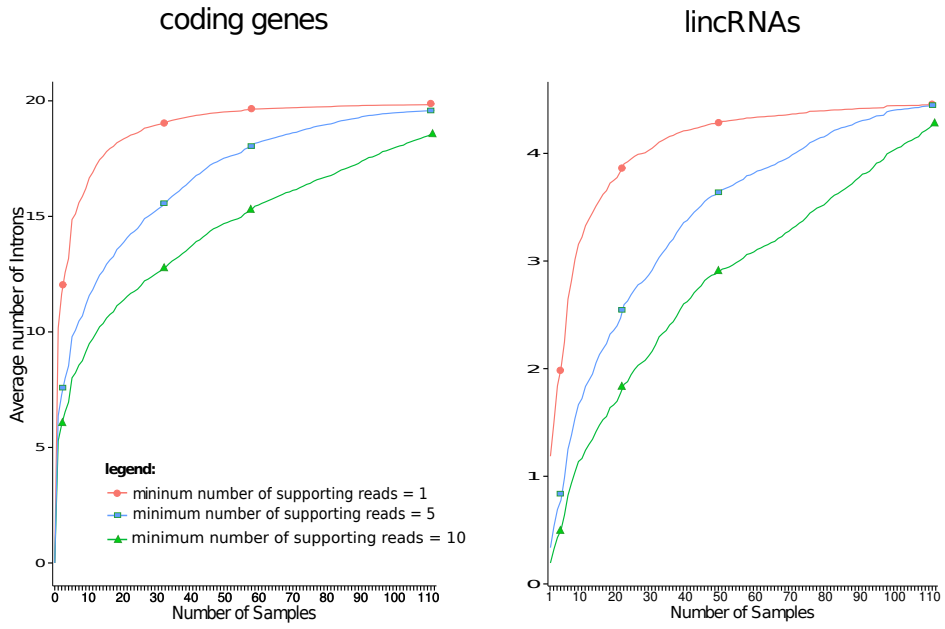



**Figure 5.2:** Saturation curves for introns. This plot shows the saturation curves for the number introns as a function of the number of independent transcriptome samples. The lncRNAs data refer to the 1,441 annotated genes in the lymphome dataset with at least one intron.

In Figure 5.3, the data was compared in more detail with the GENCODE v.19 annotation, which was used here as the reference annotation dataset, since the genes in the lymphome dataset had the largest overlap with this version. Alternative splicing is a very common phenomenon throughout the human genome[8;276] and expression levels of

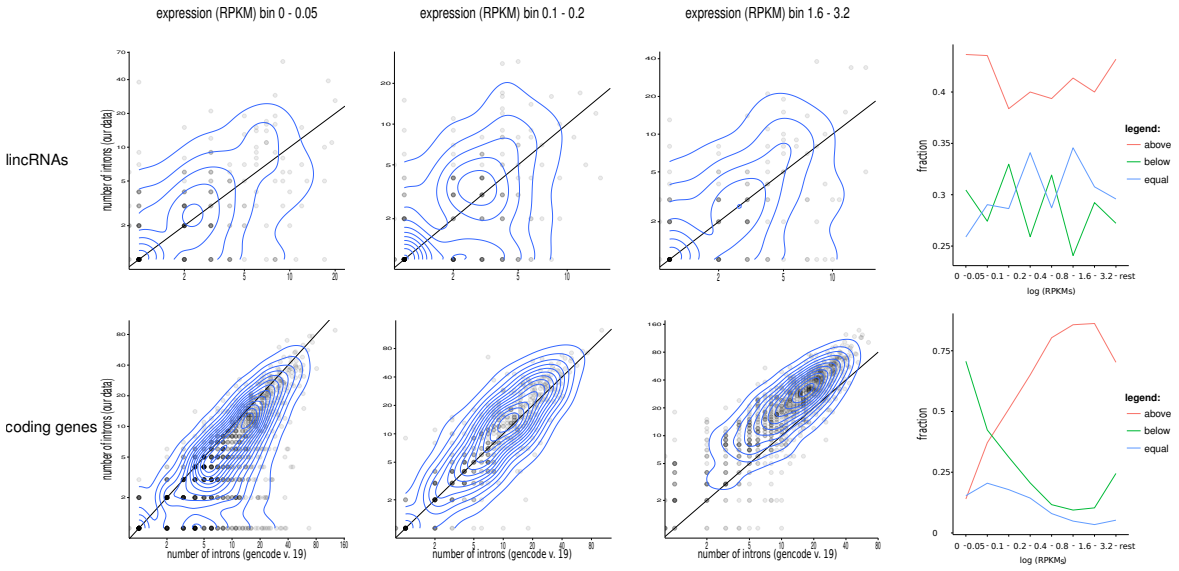
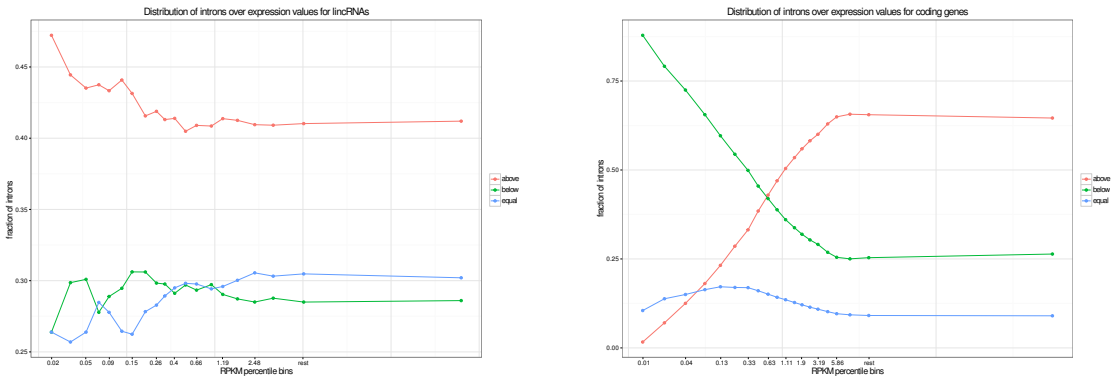



**Figure 5.3:** Scatterplots comparing lincRNAs and protein-coding genes. Scatterplots for different number of expression bins for lincRNAs and protein-coding genes. The diagonal, where $x = y$ is marked by a line. Points above the line are those genes for which more introns compared to GENCODE v.19 were calculated. Only genes with at least one intron supported by at least 10 reads were considered here. The right-most panels display the fraction of genes that show more (red), the same (blue), or fewer (green) distinct splice junctions in the lymphoma data compared to GENCODE v.19. For the coding genes there is clear dependence of these fractions on the expression level: for highly expressed mRNAs, more (rare) splice variants were systematically predicted. For mRNAs that were very lowly expressed in the lymphoma data set, GENCODE v.19 had more complex gene models. Overall, the lymphome dataset contained more introns than annotated (Wilcoxon test $p < 4 \times 10^{-10}$). In contrast, more introns in lincRNAs than annotated by GENCODE (Wilcoxon test $p < 3 \times 10^{-16}$) independent of the expression level were observed.

alternatively spliced isoforms are usually tissue specific. It is not surprising that genes with multiple exons are more likely to have alternate splice sites[23;160]. In the case of protein-coding loci, some of the splice junctions inadvertently went undetected at those loci as they were extremely lowly expressed in the lymphome transcriptomes. This was not surprising, as rare variants, of course, are easier to detect in transcriptomes where they are more highly expressed; after all, the GENCODE annotation is a composite of vastly diverse cell types and tissues. It is interesting to note, however, that more introns were observed systematically at moderate RPKM values even from the very narrowly defined cell types used here. This attests that large numbers of well defined but rare isoforms so far had eluded annotation.

Figure 5.4 shows an alternative presentation of the right-most panels showing data binned in 5-percentiles.

On comparison of the RNA-Seq data gleaned from the lymphome samples to existing GENCODE annotations it was observed that lncRNAs exhibited systematically larger numbers of exons and introns. However, the discrepancy was found to be moderate and applicable in particular to lncRNAs that already have a large number of exons annotated. Around 41% of genes were found to have more introns. 14% of the genes had at least one intron more and 19% more than two.




lincRNAs coding RNAs

**Figure 5.4:** Bins of RPKM. Normalised mean expression values quantified as reads per kilobase and million reads (RPKM). Fraction of genes with more, the same number, of fewer introns detected in the lymphoma dataset compared to GENCODE annotation. Each of the 20 datapoints represent a 5-percentile, located at the average expression value in the bin.

Furthermore, there were certain striking aspects in the lymphome data that was observed. In Figure 5.5, two examples that appear substantially more complicated in the data than in the GENCODE annotation are shown. No functional annotation was available at the moment for either locus, however, ENSG00000263470 appeared to be annotated as part of *RGS9* gene in the then immediate future version of GENCODE, version 25. As the figure shows, at least some of the additional exons also appeared in EST data tracks provided by the UCSC genome browser.

At the end of this phase, it could be observed that human transcriptome data harbour a large number of rare exons (and thus also introns) that have remained unannotated. Knowledge about the human transcriptome, especially about the lncRNAs, remains inadequate. LncRNAs cannot be called transcriptional noise anymore, in light of recent discoveries. They have been implicated in several different, in some cases non-overlapping, biological roles. However, their functions still remain poorly understood. On the other hand, functions of smaller RNAs are much better understood. Hence, the possible link between functions of small RNAs and lncRNAs, for whom lncRNAs serve as precursors, was explored. Machine learning techniques were employed for the analysis. In the next chapter, a brief outlook on the machine learning landscape in lncRNA bioinformatics will be given first.

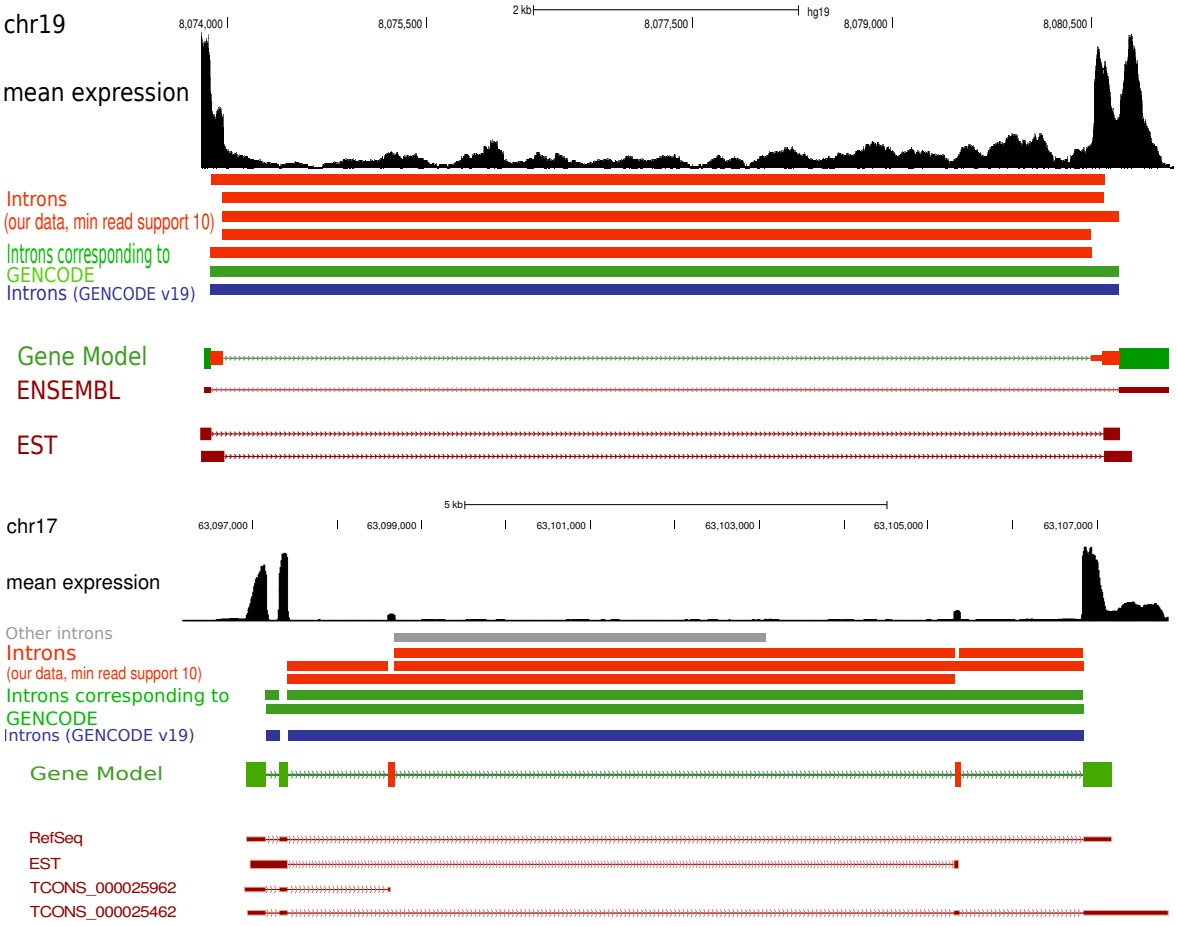



**Figure 5.5:** Examples of unannotated exons. Two examples with previously unannotated splice junctions and introns. *(top)* In ENSG00000267939 six introns and two additional exons were found compared to a single intron described in GENCODE v19. *(below)* For ENSG00000263470 eight introns plus a likely false positive compared to two introns in GENCODE was detected.

# 6

# Machine learning to detect long non-coding RNAs

This chapter now introduces the results of a comprehensive literature review, which is currently still in progress. It is concerned with the collection and evaluation of commonly used features for non-coding RNA classification and existing algorithms and tools dedicated to this task.

## 6.1 Computational techniques on the rise

Since the onset of more data storage and computing capabilities, computational biology has been developing new techniques to understand more about the information hidden in genetic codes of organisms. Focus has also been on the human genome, which has been fully sequenced[101]. Researchers have more access to transcriptomic data that are being used to unveil new elements and their potential functions computationally. Much research is being into discovering potential functions and analysing existing functions of long non-coding RNAs (lncRNAs) and it has been shown that lncRNAs play vital roles in various biological processes. However, all of the transcriptome is not yet revealed. Computational techniques are continuously being developed to study the transcriptome[277]. The sequences tend to have certain characteristics, or features, that become useful in detection technologies. In a preamble to the next chapter, here, a few of the features and established tools using machine learning for lncRNA detection in human using those features will be briefly described.

LncRNAs play important roles in many biological processes as already discussed in the introduction[1;278]. They have been observed in playing a critical role in X inactivation[150] and in various types of cancer. Although there are similarities in

sequence length between lncRNAs and mRNAs, lncRNAs function very differently. Their gene structures are also similar: lncRNAs have a similar distribution of introns and exons, but lack ORF[155]. All these similarities contribute to the difficulty in detection of and distinction between protein-coding RNAs and lncRNAs. Experimental procedures are simply too time consuming. Apart from the fact that due to the structural similarities between lncRNAs and mRNAs, there are other difficulties as well in classifying lncRNAs, or ncRNAs as a whole. Presence of genome annotation is required for most of the current approaches, which indicates that distinction of ncRNAs from protein-coding RNAs is almost entirely concentrated on species that have been well researched. Protein-coding transcripts can either be full-length or partial-length. Non-coding transcripts (which can also be truncated) parallel truncated protein-coding transcripts, causing issues in classification. As several transcript characteristics are not the same across species, building a universal classifier would not be the most efficient approach. Furthermore, accurate classification of non-coding transcripts is still missing from most existing approaches, as they display high false positive rates. Nevertheless, machine learning approaches have made some breakthroughs[279]. The possibility of bioinformatics analyses of transcriptomic data of the human genome have enabled researchers to develop machine learning inspired approaches to identify lncRNAs.

## 6.2 Commonly used features in existing approaches

The existing approaches to detect ncRNAs, specifically lncRNAs, rely on a few different features, which are mostly based upon the theory of calculating the coding potential of the transcripts. The value of the coding potential for them is far less than mRNAs, which is why this is an easy indicator adopted to distinguish the two classes.

- *Fickett TESTCODE:* It was the first method proposed to find a distinguishing factor between the two classes of RNA[280]. To circumvent the problem of detectability of initiation signals in a sequence, the authors devised a test which would enable them to identify whether a DNA sequence is coding or non-coding. It is based upon the asymmetric distribution of codons. They argued that in protein-coding sequences, the bases found in identical codon positions are the same, which is non-existent in non-coding sequences. The based their test on eight different parameters, where the first four parameters are a measure of the bases A, T, G, C, which calculates the probability of one of them being favoured in one of the three codon positions. The rest of the parameters are percentages of the contents of the bases in the sequence. Attaching weights to the parameters the coding potential of the codons is computed, which becomes the TESTCODE, or Fickett score. Inspired from the detection tool CPAT[281], the technique is presented below:

$$Fickett\,score = \sum_{i=1}^{8} p_i w_i. \tag{6.1}$$

  $p_i$ for every base $i$ is the probability of a base being favoured at a certain position

and is derived from:

$$A_1 = \text{Number of As in position 0, 3, 6,...}$$
$$A_2 = \text{Number of As in position 1, 4, 7,...}$$
$$A_3 = \text{Number of As in position 2, 5, 8,...}$$
$$A_{pos} = \frac{\max(A_1, A_2, A_3)}{\min(A_1, A_2, A_3) + 1}.$$

The positions of the nucleotides (*i. e.* C, G, T) are calculated similarly and coupled with percentage of composition of each nucleotide the values are converted into probabilities ($p$) using a lookup table in[280]. $w$ in Equation 6.1 is a weight multiplied to the probability denoting the rate of predicting coding potential by that parameter independently. This feature is capable of achieving 94% and 97% sensitivity and specificity, respectively, on lncRNA sequences, with being undecided for 18% of the sequences[281].

- *ORF length:* An *open reading frame* is a section of the DNA consisting of codons. To transcribe the DNA into mRNAs and form proteins, in case of protein-coding genes, the ribosome `reads` the ORF, commencing at the start codon (`ATG`) and ending at the stop codon (`TAA/TGA/TAG`). Ideally, reading frames can exist on both the sense and the antisense strands. The ORF is a portion of those reading frames which might have the potential to be translated. The length of the ORF is a feature used to predict if a sequence has any coding potential[282].

- *ORF coverage:* This feature is simply the ratio between the ORF length and the length of the sequence in consideration. Long ORFs are generally considered as an indicator of a coding sequence. Following that logic, if the coverage ratio is low, it probably is a non-coding sequence[283].

- *Hexamer score:* The hexamer usage bias in a given sequence is shown by this feature. Generally, coding sequences can be determined through a positive score, whereas non-coding sequences generate a negative score[284]. Adjacent amino acids are dependent on each other in a protein[281]. To exploit that property, hexamer usage bias is normally calculated. There are several different ways of determining the hexamer score. The strategy used in the tool CPAT[281] was computing the log-likelihood ratio between coding and non-coding sequences. The score was calculated for a sequence S = $H_1, H_2, ..., H_m$ with $m$ hexamers as:

$$hex_score = \frac{1}{m} \sum_{i=1} m \log(\frac{F(H_i)}{F'(H_i)}),$$

where $F(H_i)$ and $F'(H_i)$ represent the probability of each hexamer to be coding in protein-coding and non-coding sequences, respectively, with $i \, \epsilon \, (0, ..., 4095)$, the total number of hexamers possible.

- *Euclidean and logarithmic distance:* The distance of the sequence from coding and non-coding sequences is calculated. The ratio of the distances constitute the features and is explored in the LncFinder tool[285].

- *GC content:* This feature computes the ratio of total number of purine bases (either G or C) in the sequence against the length of the sequence. Higher GC content is associated to coding sequences[286].

- k-*mer:* $k$-mers are nucleotide sequences found or arranged within a DNA sequence. They denote the relative frequency of oligonucleotides. $k$ refers to the number of nucleotides or the size of the oligonucleotides in the sequence. A single base is a 1-mer, whereas the codons, which consists of three bases, are 3-mers. The information that $k$-mers encode is distributed based on the value of $k$. If $k$ is lower, the $k$-mers are more abundant and overlapping probabilities are higher than for $k$-mers with higher $k$. A 7-mer would encode more information than a 3-mer, for example, as the probability of occurrence of a particular 7-mer is much lower than that of a 3-mer. Higher $k$-mers are also computationally expensive[287;288]. The possible number of $k$-mers can be seen in Table 6.1.

**Table 6.1:** Possible number of $k$-mers. The total number of possible $k$-mers is $4^k$, but they get computationally expensive as $k$ increases.

| $k$-mer | Possible counts |
|---|---|
| 2-mer | 16 |
| 3-mer | 64 |
| 4-mer | 256 |
| 5-mer | 1,024 |
| 6-mer | 4,096 |
| 7-mer | 16,384 |

## 6.3 lncRNA detection strategies

Most of the work that has been done until now were primarily driven by the attempt to distinguish protein-coding sequences from non-coding sequences. In the majority of the instances, the problem was defined as a binary classification task, where various sequence intrinsic features were engineered and employed to separate the two classes based on the coding potential of the sequences[289]. An abundance of the usage of ORF related features can be found as those can effectively distinguish protein-coding transcripts from non-coding transcripts. Shorter ORF length and lower ORF coverage normally signify lncRNA transcripts[121;290].

### CONC

*Coding Or Non-Coding*, or *CONC* in short,[291] is one of the earliest tools to utilize these methods and regularly used as a benchmark in the successive tools developed. One of its focus areas was the distinction of the subtype of long non-coding transcripts from protein-coding transcripts. It was based on `support vector machines`[228] incorporating a plethora of features visible in a transcript coding for protein. The selected features consisted of i) peptide length (four variables): 20, 40, 80, $\geq$ 80 length intervals were selected, ii) amino acid composition (20 variables), iii) predicted secondary structure content (three variables), iv) mean hydrophobicity (one variable), v) percentage of residues exposed to solvent (one variable), vi) sequence compositional entropy (one variable), vii) number of homologues (one variable), and viii) alignment entropy (one

variable). The authors also computed $k$-mers with $k$ values being (1,2,3) and had a final feature set consisting of 180 features. The `SVMs` trained on radial basis function kernel and the hyperparameters were optimised by training on a subset of training data. The model reached 97% and 95% accuracy values for protein-coding and non-coding sequences, respectively, on 10-fold cross-validation.

## CPC, CPC2

*Coding Potential Calculator*, usually known as CPC,[283] is another early tool to implement a `support vector machines (SVM)` based classifier. Six features based upon the ORF had been incorporated into the tool. They include quality of the ORF (the higher the better) and ORF coverage. If the ORF contained a start and a stop codon, ORF integrity was another feature incorporated. The authors argued that as protein-coding transcripts are likely to have more interactions with proteins as opposed to non-coding sequences, the number of interactions would be a necessary feature. They developed a method to calculate the integrity of the protein-transcript interactions, mentioning that the higher the integrity, the more likely the transcript would be coding. To compute all the features, they used the tool *framefinder*[292] which enables the user to execute a search based on three frames and has high rates of ORF detection. For every frame, the number of hits a transcript had against known proteins was calculated by performing BLASTX[293] on the protein database UniProt Reference Clusters[294]. The quality of the hits was another feature that was considered, going by the fact that coding transcripts have higher quality hits. This was computed using:

$$S_i = mean_j\{-log_{10}E_{ij}\}[i\epsilon\{0,1,2\}] \tag{6.2}$$

$$hit\ score = mean_{i\epsilon\{0,1,2\}}\{S_i\} = \frac{\sum_{i=0}^{2} S_i}{3}$$

In 6.2, $E_{ij}$ stands for the $E$-value (for BLASTX queries) of $j$th high scoring segment in frame $i$, $S_i$ is the average quality of the segment, and *hit score* gives the mean $S_i$ across three frames. The concentration of hits among the three frames used by *framefinder* was the third feature extracted: coding transcripts would have the hits concentrated in a single frame in contrast to non-coding transcripts. *frame score* is actually the variance of $S_i$:

$$frame\ score = \frac{\Sigma_{i=0}^{2}(S_i - \overline{S})^2}{2}.$$

If the score is high, then the transcript being protein-coding rises. An `SVM` based model was trained using a standard radial basis function kernel, whose $C$ and *gamma* parameters were determined through a grid search. CPC reported an accuracy of 96% based on 10-fold cross-validation tested on two datasets containing non-coding transcripts and one containing protein-coding transcripts. On lncRNAs particularly, accuracy levels were reported to be around 76%.
An upgrade to CPC is the *CPC2* tool[295]. It is also based on a support vector machine and uses four features as opposed to six in the original tool. Implementing a `random`

`forest` model with recursive feature elimination technique on a collection of sequence intrinsic features the authors have derived Fickett score, ORF length, ORF integrity (presence of both start and stop codons if a ORF was present) and isoelectric point of a predicted peptide as the four main features to be fed into the `SVM` for classification. They reported CPC2 to not only be faster (the authors reported to be 1000x faster), but also more accurate than its predecessor. As opposed to CPC's accuracy rate of 76% for lncRNAs, CPC2 reported 94% accuracy. It is also more consistent in classification accuracy rates than CPC, the authors wrote.

## CNCI, CNIT

An annotation-free classification tool from cross-species transcriptomes relying upon sequence intrinsic composition is called *Coding-Non-Coding Index* or *CNCI*[296]. The authors achieved that using `support vector machines` with a standard radial basis function kernel trained on five features. They identified the coding domain sequence (CDS) by employing a system to analyze nucleotide triplets (essentially 3-mers), which harks to the hexamer usage bias feature. Employing the usage frequency of each possible nucleotide triplets, a 64*64 matrix was created, as described below:

$$X_iN = \sum_{j=1}^{n} S_j(X_i).$$

$S_j(X_i)$ is the number of times a triplet $X$ occurs in the sequence $i$. $T$ is frequency of all triplets given by:

$$T = \sum_{i=1}^{m} X_iN.$$

$m$ is $64 * 64$, the number of adjacent triplets that can be generated.
$X_iF$ denotes usage frequency of the triplets:

$$X_iF = \frac{X_iN}{T}.$$

The authors used a sliding window of size 150 nucleotides to traverse each transcript and generated six reading frames from which they calculated the CDS most likely to be transcribed implementing maximum interval sum function[296]. The length and the quality of the most suitable CDS ($S$-score) were used as two main features fed into the `SVM`. Two more features were computed concerning the lengths and scores (qualities) of all the coding domain sequences detected as shown below:

$$length - per = \frac{M1}{\sum_{i=0}^{n}(Y_i)} \; i \, \epsilon \, (1, ..., 6)$$
$$score - dist = \frac{\sum_{j=0}^{n}(S - E_j)}{5} \; j \, \epsilon \, (1, ..., 5)$$

$M1$ denotes the length of the best CDS calculated based upon the $S$-score, $Y_i$ the length of each frame, $E_j$ scores of the other five frames.
Although these features were capable of distinguishing protein-coding sequences from

non-coding sequences, another feature category was added which indicated the coding bias of the 61 codons, without the stop codons, detected in the CDS. The SVMs were trained on standard radial basis function kernel with the hyperparameters set by default. The tool achieved around 97% accuracy on 10-fold cross-validation when trained on human protein-coding sequences and long non-coding sequences. It also achieved 94% accuracy on unseen human test data. Despite being trained on human transcripts, CNCI was reported to separate protein-coding and long non-coding transcripts fairly well across other species including mouse, *Caenorhabditis elegans* and orangutan.
*Coding-Non-Coding Identifying Tool*, or *CNIT*[297], is the successor to the above mentioned tool and is reported to be 200x faster. It consists of the same set of features as CNCI and is trained on an `SVM`, too, in particular on the `XGBoost` software implementation[298]. The model was trained on both human and *Arabidopsis thaliana* protein-coding and non-coding transcripts and tested on 10 animal and 26 plant species. On human test sequences, it achieved 98% accuracy levels.

## PLEK

*PLEK*[287] is another alignment-free tool designed to distinguish between mRNAs and lncRNAs utilising `SVMs` with a radial basis function kernel, whose $C$ and *gamma* parameters were computed through a 10-fold cross-validation grid search. Developed to deal with classification errors, specially in case of *de novo* transcripts, where there might be indels, PLEK implemented $k$-mer frequencies as a feature. Choosing $k$ values from 1 to 5, the authors computed a weight function for each $k$-mer length with a sliding window approach to have a relationship between a $k$-mer and the sliding window. The total number of $k$-mer patterns amounted to 1,364.

$$
\begin{aligned}
s_k &= l - k + 1 \\
w_k &= \frac{1}{4^{5-k}} \\
f_i &= \frac{c_i}{s_k} w_k \quad [k = 1, ..., 5;\ i = 1, ..., 1364]
\end{aligned}
$$

The above set of equations outlines the process of utilising the $k$-mer information. For a transcript of length $l$, a sliding of length $k$ was used, with $c_i$ being incremented by one if a pattern $i$ was matched by a string inside the window. $s_k$ stored the number of times the window could slide.
The $k$-mer usage frequencies were normalized using the libsvm package[229] and used as features for the SVM. On training the SVM in capacity of a binary classifier on human protein-coding and long non-coding sequences, the tool achieved more than 95% accuracy on 10-fold cross-validation. Cross-species validation returned similar accuracy levels as CNCI on same unseen data. On a separate set of test data containing *de novo* transcripts, the tool reported around over 93% accuracy levels.

## LncFinder

*LncFinder* is a tool to detect as well as predict novel lncRNAs[285]. The authors behind this tool exploited three different feature categories. Apart from intrinsic sequence composition, which contained features to quantify hexamer usage bias using distances, secondary structure features and physiochemical properties of the sequences were used. Each feature set contained three features, except secondary structure features, which contained four. The minimum free energy for the structure features was calculated using the RNAfold program of the ViennaRNA package[299]. Abandoning the $k$-mer scheme implemented in some other tools, the authors proposed two feature categories to quantify hexamer usage bias. Each category has three features: genomic distance to lncRNA, genomic distance to protein-coding transcript and distance ratio.

$$
\begin{aligned}
distlnc &= \sqrt{\sum (freq_{seq}(i) - freq_{lnc}(i)^2} \\
log(distlnc) &= \frac{1}{n} \sum \ln \frac{freq_{seq}(i)}{freq_{lnc}(i)}, \quad i = 1, ..., 4^k \\
log(distratio) &= \frac{\log distlnc}{\log distpct}
\end{aligned}
$$

In the above equations $freq_{seq}$ stands for $k$-mer frequency, $freq_{lnc}$ stands for the mean lncRNA $k$-mer frequency, $i$ stands for different types of $k$-mers, $n$ denotes the total number of $k$-mers. For protein-coding sequences, *distpct* was calculated similarly.

The authors employed physiochemical properties of nucleotides as a feature set, as each nucleotide has one EIIP value[300]. They utilised fast Fourier transform (FFT) to build a power spectrum to take advantage of the $N/3$ position of a sequence, which shows a peak for protein-coding sequences, but not for lncRNA sequences. Through that strategy they compiled three features, namely the signal at the $N/3$rd position, the signal-to-noise ratio and the average power for all the sequences.

The authors then trained various models, namely, `logistic regression, SVMs, random forest, ELM, deep learning` and selected `SVM` over the others based on 10-fold cross-validation, which reported above 96% in accuracy trained on human transcripts.

## iSeeRNA

One more `SVM` based identification tool is *iSeeRNA*[301]. This tool leverages three feature categories comprising conservation scores, ORF details and nucleotide sequences. Mean of the phastCons scores[302;303] of every nucleotide for a sequence was calculated. ORF length and ORF coverage were the other two features computed. The last feature category constitute two 2-mers and five 3-mers: [GC, CT, TAG, TGT, ACG, TCG](with the fifth 3-mer not mentioned). iSeeRNA was trained on these 10 features using an SVM on a standard radial basis function kernel with optimized hyperparameters ($C$ and *gamma*). Training it on human and mouse coding and non-coding sequences and

reported 95.4% and 94.2% accuracy values, respectively, on 10-fold cross-validation. It also returned detection accuracy rates of 96.1% and 94.7% on human test dataset comprising lncRNAs and mRNAs, respectively. A collection of *de novo* lncRNA transcripts was predicted correctly with more than 97% accuracy rate.

## lncRScan-SVM

Yet another `SVM` based tool is called *lncRScan-SVM*[304]. It mainly depends on six features comprising transcript length, exon count of a gene, and mean exon length, mean conservation score, likelihood of a codon sequence in a sequence of nucleotides (*txCdsPredict* from *UCSC Table Browser*), and the standard deviation of stop codon counts between three translated frames. The final feature was chosen on the basis of standard deviation of stop codon counts of lncRNA transcripts is more than protein-coding transcripts. It was formulated as:

$$SCS = \sqrt{\frac{1}{3}\sum_{i=0}^{2}(SCC_i - \overline{x})^2},$$

where $\overline{x} = \frac{1}{3}\sum_{i=0}^{2} SCC_i$ and is the mean of stop codon counts of three frames ($SCC_0$, $SCC_1$, $SCC_2$).
By training the model on the human transcripts, the tool reported around 91% accuracy values. On the test dataset, it performed similarly.

## CPAT

Another tool that was published around the same time as CNCI was *Coding Potential Assessment Tool*, or *CPAT*[281] It is an alignment-free tool based on a `logistic regression model`. The four features that it was built on comprised the ORF size, ORF coverage, Fickett score and hexamer usage bias. The Fickett score is calculated as a probability of a nucleotide being favoured in its position in a codon. Adjacent amino acids are dependent on each other in a protein. To exploit the property of adjacent amino acids being dependent in a protein and not so much in a non-coding RNA the hexamer usage bias was normally calculated. Training CPAT on human coding and non-coding transcripts generated over 99% accuracy levels on 10-fold cross-validation. The tool was tested on an unseen dataset with 96% and 97% of sensitivity and specificity, respectively.

## FEELnc

Based on `random forest` model, another alignment-free tool to identify lncRNAs from mRNAs is *FlExible Extraction of LncRNAs*, or *FEELnc*[305]. It also possessed the ability to identify potential lncRNA biotypes and annotate them. The tool computed ORFs and annotated them in five different categories: from a *strict* mode, where the

ORF contains both the start and stop codon, to a *relaxed* mode, where the whole sequence was considered. $k$-mer frequencies ($k$ ranging from 1 to 12) for each mRNA ORF sequence and lncRNA sequence were calculated and a score was assigned to each transcript for each $k$-mer size. Using a `random forest` implementation the tool determined and optimized coding potential score for every sequence. The sequence would be long non-coding if the score tended to 0, coding if it was closer to 1. The other features used were ORF length, ORF coverage and sequence length. Human and mouse protein-coding and long non-coding sequences were used to train the `random forest` classifier and it achieved around 91% accuracy.

## LncRNA-ID

*LncRNA-ID* is another `random forest` based identification tool designed to identify lncRNA sequences among a set of protein-coding and long non-coding sequences[306]. The tool employed 11 features consisting of ORF related features, ribosomal interaction related features (when ribosomes interact with mRNAs during protein translation), and protein conservation scores using profile `hidden Markov model` based alignment. To compute ribosomal interaction related features, the first thing the authors considered was the Kozak motif: GCCRCCAUGG, as nearly all ribosomes interact with AUG[307;308]. The ribosomal coverage was calculated as the change in binding energy:

$$rc = \sum_{i=1}^{L} \{N_i | \delta < 0\},$$

where $\delta_i$ is the free energy at position $i$, $L$ the sequence length, and $N_i$ the number of base pairs starting at $i$. It was expected that coding transcripts had more ribosomal coverage.
The relative degree of ribosome occupancy bias at the terminal binding site of a sequence was calculated as:

$$rrs = \frac{\frac{\text{rc of ORF}}{\text{length of ORF}}}{\frac{\text{rc of 3'UTR}}{\text{length of 3'UTR}}}.$$

This occupancy bias is normally larger for coding transcripts.

The accuracy levels reached 96% when the `random forest` classifier was trained on these features on a human transcripts dataset. The authors tested the classifier on the dataset used by CPAT and reached accuracy rates close to 95%.

## LncRNApred

A `random forest` based model was the preferred classifier for the developers of the tool *LncRNApred*[309]. The authors retrieved human mRNA and lncRNA transcripts and subjected those transcripts to *self organizing feature map* or `SOM` clustering to select representative samples for the training dataset. The authors initially chose 89 features, but after feature selection the number was reduced to 30. One of the features used was called signal-to-noise ratio which essentially converted a sequence into four binary

sequences according to the nucleotides. For example, a sequence `TAGGTCAT` would be encoded as:

$$u_A = \{0,1,0,0,0,0,1,0\} \qquad u_T = \{1,0,0,0,1,0,0,1\}$$
$$u_G = \{0,0,1,1,0,0,0,0\} \qquad u_C = \{0,0,0,0,0,1,0,0\}$$

where $u_b$ denotes the representation of the sequence to be encoded for each nucleotide $b$; $b \epsilon$ {A,T,G,C}. For every binary sequence, a complex sequence was created through *Discrete Fourier Transform* and combining all four, a resultant power spectrum was achieved. As nucleotide usage bias is profoundly observed in coding transcripts, as opposed to non-coding transcipts where nucleotides are evenly distributed across codons[309], the power spectrum at 1/3 length of every sequence was incurred as an important feature by the authors, as a clear distinction between coding and lncRNA transcripts was visible. The other features were two ORF related features, $k$-mer features ($k$ ranging from 1 to 3), GC content and sequence length. The random forest model trained on these features resulted in 93% accuracy rates for human.

## lncRNAnet

lncRNAnet is a deep learning based approach to detect lncRNAs[310]. The authors implemented recurrent neural networks (`RNN`) to determine intrinsic features of lncRNAs and convolutional neural networks (`CNN`) to identify stop codons in ORFs. RNN follows the nature of an acyclic graph: it learns the sequential behaviour of the data by feeding the output of the previous cell into the next one. The parameters of each cell are carried forward and the network behaves flexibly, which is why it is used for text classification in sequence-to-sequence learning. CNN has been very effective in the branch of image classification. It implements lots of convolution filters or layers to extract connections from sparse data and reduces the number of parameters as compared to other neural networks[311]. In lncRNAnet CNN has been applied to detect the region between two stop codons, as ORF detection can be tough if there are the start codons are non-canonical, the authors argued. On identification of all the stop codons, a ORF indicator was populated with values 0 and 1 for each nucleotide's presence in a ORF, with the aim to find the longest ORF. All sequences smaller than the largest sequence were padded to match the maximum length and encoded as four-dimensional tensors. Each nucleotide was encoded as: A: [1,0,0,0], C: [0,1,0,0], G: [0,0,1,0], T: [0,0,0,1]. The ORF related features were projected as two-dimensional tensors. The transformed protein-coding and lncRNA transcripts, along with the ORF indicator, were passed through an RNN with one stacked layer and the output of the network was passed through a two-dimensional fully connected layer. The data was taken from GENCODE v.25, split 85-15, and trained using 100 hidden layers for 200 epochs, incorporating features namely sequence, ORF indicator, ORF length and ORF coverage, which produced 99% accuracy based on 5-fold cross-validation. The authors maintained that ORF indicator is a key feature in lncRNAnet, which also enabled them to successfully identify lncRNAs, irrespective of their sequence lengths.

## lncADeep

Another deep learning based approach is *lncADeep*, whose purpose is to identify novel lncRNAs and functionally annotate them. This tool operates free of any reference based on a deep belief network (DBN)[312]. The tool deals with both i) full length and ii) full and partial length mRNA transcripts and combines them with lncRNA transcripts. The features incorporated are ORF length, ORF coverage, EDP of ORF[313], hexamer score, UTR coverage, GC content, Fickett score, HMMER index, and longest CDS. The authors implemented a DBN from restricted Boltzmann machines (RBM) for identification of lncRNAs. lncADeep also deals with lncRNA-protein interaction to identify lncRNA functions. Structural features like folding energy and hydrogen bonding are used besides sequence features to construct a deep neural network (DNN) to predict interactions. The tool achieved 98% sensitivity on 10-fold cross-validation when identifying lncRNA transcripts on a training set based on full length mRNA transcripts. On training the tool on partial length mRNA transcripts, it achieved 94% sensitivity while identifying lncRNAs from a dataset comprising both full and partial length and only partial length mRNA transcripts, respectively. It claims an accuracy of 97% in lncRNA identification.

## DeepLNC

*DeepLNC* is another tool based on deep neural networks (DNN)[314]. The only feature set used here is $k$-mer frequency combinations for $k$ values ranging from 2 until 5. The counts were normalised using Shannon entropy. The authors implemented a binary classification model based on DNN to separate mRNA transcripts from lncRNA transcripts. The tool reached accuracy levels of 98% on 10-fold cross-validation with $k$-mer combinations of [2,3,5].

The described tools constitute several different characteristics that make them stand out from each other. Some of the characteristics which make them interesting are compiled in Table 6.4.

## 6.4 Tabular overview of the tools

**Table 6.2:** LncRNA detection tools. Some of the essential characteristics of the lncRNA detection tools mentioned are summarised.

| Tool | Algorithm | Species | Features | Performance |
|---|---|---|---|---|
| **CONC** | SVM | Eukaryotes (both protein-coding and non-coding genes) | peptide length, amino acid composition, predicted secondary structure content, mean hydrophobicity, percentage of residues exposed to solvent, sequence compositional entropy, number of homologues, alignment entropy | 10-fold CV on protein-coding: F1-score: 97.4% ⋆ Precision: 97.1% ⋆ Recall: 97.8% ◙ On non-coding: F1-score: 94.5% ⋆ Precision: 95.2% ⋆ Recall: 93.8% |
| **CPC** | SVM | Eukaryotes (both protein-coding and non-coding genes) | ORF features (quality, coverage, integrity), number of BLASTX hits, hit score, frame score | 10-fold CV: 95.77% ⋆ Accuracy on Rfam database (non-coding): 98.62% ⋆ RNADB (non-coding): 91.5% ⋆ EMBL cds (protein-coding): 99.08% ◙ Accuracy in lncRNA detection: 76.2% |
| **CPC2** | SVM | Species neutral, trained and tested on animals and plants (both protein-coding and non-coding genes) | ORF features (quality, coverage, integrity), Fickett score, isoelectric point | Accuracy: 96.1% ⋆ Specificity: 97% ⋆ Recall: 95.2% ◙ Accuracy in lncRNA detection: 94.2% |
| **CNCI** | SVM | Vertebrates, plants, orangutan | Adjacent nucleotide triplets, sequence score, codon-bias, most-like CDS (MLCDS), length-percentage, score-distance | 10-fold CV accuracy on human: 97.3% ◙ Minimum average error for vertebrates $< 0.1$ ⋆ Plants: 0.24 |
| **CNIT** | SVM | 11 animal species, 26 plant species | max_score of MLCDS, standard deviation of MLCDS scores and MLCDS lengths, frequency of 64 codons | Accuracy on human: 98% ⋆ Mouse: 95% ⋆ Zebrafish: 93% ⋆ Fruit fly: 93% ⋆ *A. thaliana*: 98% |
| **PLEK** | SVM | 11 vertebrates | $k$-mer frequency (for $k$=[1,5]) | 10-fold CV accuracy: 95.6% |
| **LncFinder** | SVM | Trained on human, tested on human, mouse, wheat, zebrafish, chicken | genomic distance to lncRNA, genomic distance to protein-coding transcript, distance ratio, EIIP value | 10-fold CV accuracy: 96.87% |

| Tool | Algorithm | Species | Features | Performance |
|---|---|---|---|---|
| **iSeeRNA** | SVM | human, mouse | frequency of six $k$-mers (GC, CT, TAG, TGT, ACG, TCG), conservation score, ORF length and proportion | Accuracy in human lncRNA detection: 96.1% ✡ Mouse: 94.2% ◙ Accuracy in human protein-coding gene detection: 94.7% ✡ Mouse: 92.7% |
| **lncRScan-SVM** | SVM | human, mouse | sum of lengths of exons, frequency of exons, mean exon length, standard deviation of stop codon frequency, txCdsPredict | Two test sets created based on i) random protein-coding and lncRNA sequences and ii) only dissimilar sequences. Accuracy on set A for human: 91.54% ✡ Mouse: 92.21% ◙ On set B for human: 91.45% ✡ Mouse: 92.2% ◙ MCC on set A for human: 83.17% ✡ Mouse: 84.59% ◙ On set B for human: 82.99% ✡ Mouse: 84.69% ◙ AUC on set A for human: 96.39% ✡ Mouse: 96.62% ◙ On set B for human: 96.39% ✡ Mouse: 96.64% ◙ |
| **CPAT** | logistic regression | human | ORF length, ORF to transcript length ratio, Fickett score, hexamer usage bias | 10-fold accuracy: 99% ✡ Precision: 96% |
| **FEELnc** | random forests | human, mouse | ORF features (coverage, length), sequence length, coding potential score, $k$-mer score based on frequency | Accuracy for human: 91.9% ✡ Mouse: 93.9% ◙ Sensitivity for human: 92.3% ✡ Mouse: 93.8% ◙ Specificity for human: 91.5% ✡ Mouse: 94.1% ◙ F score for human: 91.9% ✡ Mouse: 95.6% ◙ MCC for human: 83.8% ✡ Mouse: 85.6% |
| **LncRNA-ID** | random forests | human, mouse | ORF related features, ribosomal interaction related features, protein conservation scores | Specificity on human: 95.28% ✡ Mouse: 92.1% ◙ Recall on human: 96.28% ✡ Mouse: 94.45% ◙ Accuracy on human: 95.78% ✡ Mouse: 93.28% |
| **LncRNA-pred** | random forests | human, mouse, "other species" | ORF features, signal-to-noise ratio, $k$-mer frequency (for $k$=[1,3]), GC content, sequence length | Accuracy on human: 92.96% ✡ Mouse: 94.3% ◙ Specificity on human: 92.5% ✡ Mouse: 93.48% ◙ Recall on human: 93.42% ✡ Mouse: 95.27% ◙ Accuracy on other species for lncRNAs: 97.78% |

| Tool | Algorithm | Species | Features | Performance |
|---|---|---|---|---|
| **lncRNAnet** | convolutional neural network, recurrent neural network | human, mouse | sequence, ORF features (length, coverage, indicator) | 5-fold accuracy: 99% ◙ Accuracy on human: 91.79% ✲ Mouse: 91.83% ◙ Specificity on human: 87.66% ✲ Mouse: 89.03% ◙ Sensitivity on human: 95.91% ✲ Mouse: 94.63% ◙ AUC on human: 96.72% ✲ Mouse: 96.67% ↯Also available are test results on 11 different species and on experimental NGS data. |
| **lncADeep** | deep belief network | human, mouse | ORF features (length, coverage, hexamer score of longest ORF, entropy density profile), UTR coverage, GC content of UTRs, Fickett score, HMMER index | Precision for lncRNA detection from full-length mRNA transcripts: 97.2% ✲ Recall: 98.1% ✲ Average harmonic mean: 97.7% ◙ Precision for lncRNA detection from both full and partial-length mRNA transcripts: 94.5% ✲ Recall: 93.8% ✲ Average harmonic mean: 94.2% ◙ Precision for lncRNA detection from partial-length mRNA transcripts: 90.3% ✲ Recall: 93.8% ✲ Average harmonic mean: 92% |
| **DeepLNC** | deep neural network | human | $k$-mer combinations (for $k$=[2,5]) | 10-fold CV accuracy: 98.07% ✲ MCC: 96% ✲ Recall: 98.98% ✲ Precision: 97.14% ✲ AUC: 99.3% |

# 7

# lncRNAs playing host to smaller RNAs

Many small nucleolar RNAs (snoRNAs) and many of the hairpin precursors of microRNAs (miRNAs) are processed from the same genomic loci as long non-coding RNAs(lncRNAs). The aim of this research work was to study the relationships between the hosted and the host genes and find out, if there exists any signal to reliably distinguish the three classes of lncRNAs from each other using machine learning techniques. The research question is described first, followed by brief introductions to miRNAs and snoRNAs. Thereafter, the study in itself is described in detail.

## 7.1 The research question

A wide variety of molecular and biological functions have been reported for lncRNAs. With advancing techniques for RNA detection and prediction in the genomic space, more lncRNAs are being discovered continuously. As lncRNAs have similarities to mRNAs, for example, their length, it is not surprising that this group of RNAs is involved in an array of biological activities, such as gene expression regulation and post-transcriptional repression of other RNAs, among others. Specific lncRNAs regulate chromosome architecture and chromatin remodeling. They modulate inter- and intra-chromosomal interactions and regulate recruitment of chromatin modifiers. LncRNAs have also been found to regulate turnover, translation, and post-translational modification of mRNAs. Certain lncRNAs regulate transcription by forming R-loops, thereby recruiting transcription factors and interfering with the RNA pol-II machinery to inhibit transcription[1]. As miRNAs are transcribed from pri-miRNAs by the same enzyme, this act of interference is of particular interest. Although biologically very significant, the functions of lncRNAs do not seem to be defined by their sequence or structural characteristics. In contrast to protein-coding genes, whose functions are closely tied to the respective protein families, lncRNA functions cannot be predicted

based upon sequence similarity alone. Consequently, it has remained impossible to predict the biological function, along with the molecular mechanism of an lncRNA solely from its sequence.

It can also be seen in smaller non-coding RNAs, that they are heavily structured and are readily recognisable through their highly conserved sequences. Spliceosomal RNA or ribosomal RNA detection and function prediction depend greatly on their conserved sequences. The cloverleaf shape of tRNAs or the ultra-stable hairpins of miRNAs are examples of class specific features that are useful in detecting either of two mentioned groups of RNAs[315]. MiRNAs are classified into families depending upon who they target[316;317]. For instance, miR-141 and miR-200c are members of the miR-200 miRNA family, which have the same sequence except one nucleotide in their seed region. It was observed that following the removal of the locus of each miRNA, their targets did not overlap[318]. The miR-25/32/92/363/367 family includes an unrelated miRNA, miR-25, because it shows the same seed structure as the others and possesses common targeting preferences. However, miR-200a and miR-200b, also similar in sequence but for one nucleotide, are grouped into different families since their targeting preferences vary[317]. The cluster of miR-100:let-7:miR-125 is one the most deeply conserved clusters, whose residents are generally co-transcribed; however, unlike its partners, the miRNA let-7 is suppressed in some mammalian cancer cells[318;319].

Unsupervised clustering using normalised $k$-mer abundances as similarity measure between sequence and function of lncRNAs revealed an association of $k$-mer profiles with lncRNA function, in particular with protein binding properties and sub-cellular localisation. The authors behind this piece of work[288] designed a method, called SEEKR, to calculate $k$-mers inside a sliding window of a specified length traversing an lncRNA sequence to create a standardized length matrix of $k$-mer profiles. They defined a $z$-score for each lncRNA as:

$$z = \frac{(\text{k-mer count per kb}) - (\text{mean count per kb in group})}{\text{kmer s.d. in group}},$$

which populated the matrix. The method can quantify non-linear sequence relationships, that means using Pearson's correlation it can measure similarity and differences between sequences. *cis*-activating lncRNAs such as *HOTTIP* and *HOTAIRM1* group together due the presence of GC-rich $k$-mers, whereas *XIST* and *ANRIL*, *cis*-repressive lncRNAs, belong to a different group since they were found to be high in AU-rich $k$-mers. The authors argue that the lack of sequence similarity was not a barrier in having similar $k$-mer profiles and can be helpful in understanding relations between lncRNA genes better in terms of functions.

With a plethora of RNA-binding proteins typically recognising a wide array of local binding motifs that maybe structured, modular and gapped[320], the correlation of short $k$-mers and function is not surprising. Nevertheless, it remains unclear whether there are distinct, well-separated classes of lncRNAs or whether the universe of lncRNAs is organised as a continuum of functions and associated molecular features.

To this end, an exception to the rule are the lncRNAs whose genomic loci overlap with that of the smaller RNAs making both the groups co-dependent. These lncRNAs act as host genes for the generation and processing of miRNAs and snoRNAs, similar

to the condition when a pri-miRNA would act as a precursor to snoRNAs or vice-versa. They serve as hosts to the sponges that modify the miRNA pool by acting as decoys for miRNA as well[140]. Out of these molecules, the host genes of snoRNAs and miRNAs can be recognised easily, since the smaller RNAs lodged within them (hereafter referred to as payloads) are evolutionarily well-conserved. The snoRNA host genes (SNHGs) and the miRNA host genes (MIRHGs) undergo distinctive due to the difference in their payloads. The snoRNAs in human are exclusively located in the introns of the SNHGs (and protein-coding genes) indicating that the processing of the snoRNAs is linked to splicing[321;322]. Lykke-Andersen et al.[323] reported that nonsense-mediated RNA decay degraded almost all of the mRNA or lncRNA isoforms produced by snoRNA host genes in human. Some of the snoRNA genes maybe partially or fully present in exons of the longer RNAs, which are released through alternatively spliced isoforms and in turn lead to selective expression of the intronic snoRNAs, and the isoforms are subjected to degradation[322] (Figure 7.5).

SNHGs have been receiving growing interest particularly in cancer research. Tong et al.[324] reviewed a snoRNA host gene, *SNHG15* that is dysregulated in a wide variety of cancers including hepatocellular carcinoma, glioma, breast, pancreatic, and lung cancer, and interacts with several distinct molecular mechanisms. In almost all the cancer types it is involved in, *SNHG15* shows a marked over-expression. *SNHG20* has also been observed to be over-expressed in hepatocellular carcinoma, colorectal, bladder, and ovarian cancer[325].

The tumour suppressor gene *GAS5* could inhibit miR-182-5p and miR-221 expression, thereby suppressing colon cancer cell proliferation. It regulates miR-137 transcription leading to inhibition of cell proliferation in breast cancer and melanoma. The *GAS5* gene encodes snoRNAs like SNORD47 in its introns[139]. The gene has been suggested to possibly be functional in some other biological roles, too, for instance binding to a protein, forming a RNP complex. *GAS5* acts as a sponge for miR-23a, down-regulating it and inhibiting cardiomyocyte hypertrophy[326;327].

Most SNHGs are reported to function as miRNA sponges. Li. et al. found that *SNHG3*, which is a cancer-associated fibroblast (CAF) lncRNA, was regulated by miR-330-5p in breast cancer cells. It acts as a molecular sponge to miR-330-5p, which in turn down-regulates Pyruvate Kinase M1/M2 (PKM) expression in tumour cells, thereby enhancing breast tumour cell proliferation[328]. Some SNHGs are involved in digestive and respiratory cancers. *SNHG1* and *SNHG5*, for example, are implicated in colon cancer, gastric cancer, and liver cancer among others. *SNHG1* acts as a sponge to miR-497/miR-195-5p and could act as a great influence in colorectal cancer cell proliferation. *SNHG5* could promote liver cancer cell proliferation by up-regulating CTNNB1, MYC, and CCND1 expression, thereby activating the Wnt signaling pathway and inducing epithelial-mesenchymal transition (EMT)[329].

The secondary functions of snoRNAs seemingly are not coupled with functions of the host genes. While snoRNAs do date back to a common ancestor of Eukarya and Archaea[330], non-coding SNHGs are evolutionary much younger and seem to have appeared comparably recently in animal evolution[321]. Biological functions of SNHGs exerted by mature, spliced SNHGs have also arisen secondarily.

In contrast to SNHGs, very few MIRHGs have known functions beyond harbouring their miRNA payload. *MIR100HG* has been reported to interact with HuR/ELAV1[331] and to form RNA-DNA triplex structures with the p27 locus[281]. There is evidence that this gene might also function as an miRNA sponge[332]. *MIR31HG*, hosting the miRNA miR-31, does not actually act as a sponge to its hosted gene, nor regulates it, but acts as a sponge to tumour suppressor miR-361 and is down-regulated in osteosarcoma. It is, however, upregulated in lung cancer and colorectal cancer[333]. There are other lncRNAs, though, that regulate miRNA expression levels by acting as sponges to miRNAs[140;141]. The lncRNA *MALAT1* was found to be targeted by miR-22-3p, which in turn inhibited *MALAT1* expression levels in endothelial cells[334]. miR-675 is found embedded in the ancient lncRNA *H19* and has been observed to be expressed in placental cells. It targets IGF1R and inhibits placental growth[335].

The question this phase of the research work set out to answer was if there is any credible difference between spliced SNHGs and MIRHGs. To put it differently: are these two groups distinct classes of lncRNAs? Although they do not appear as distinct clusters in the map of lncRNA universe proposed by Kirk et al.[288], they may still be distinguished employing more complex features than $k$-mer distributions. Distinguishing lncRNAs from mRNAs computationally is a topic that has already been discussed with a lot of fanfare. Several tools have been proposed using traditional programming as well as machine learning techniques and those have already been reviewed in the previous chapter (6). MiRNAs and snoRNAs have also been shown to be easily distinguished using sequence intrinsic features employed by machine learning techniques. Tools such as miRDeep[336], MiPred[337], SnoReport[338] have been developed to predict the presence of miRNAs or snoRNAs in a particluar genomic space[339]. A more recent tool MuStARD, built deploying the deep learning method of a convolutional neural network (CNN), has a model trained on human pre-miRNAs and snoRNA precursors separately, and it can successfully identify small RNAs from a genomic region[340].

Here, the attempt was made to train a machine which could reliably separate SNHGs and MIRHGs from each other, and from a control group, a background set of lncRNAs that harbour neither snoRNAs nor miRNAs, which will be referred to as NoHGs hereafter. This is a story of deploying both supervised and unsupervised methods of machine learning upon a collection of host genes and non-host genes with the goal of achieving an optimised set of features and parameters that could perform the task successfully.

## 7.2 Small regulatory RNAs

### 7.2.1 MicroRNAs

MicroRNAs (miRNAs) are a group of small regulatory RNAs that guide post-transcriptional repression and gene expression of messenger RNAs (mRNAs) and other RNAs. MicroRNAs are part of a broader group of small RNAs, which also include small interfering RNAs (siRNAs) and piwi-interacting RNAs (piRNAs), that inhibit functions and curb undesirable transcripts[341]. The mature RNAs are 22 nucleotides

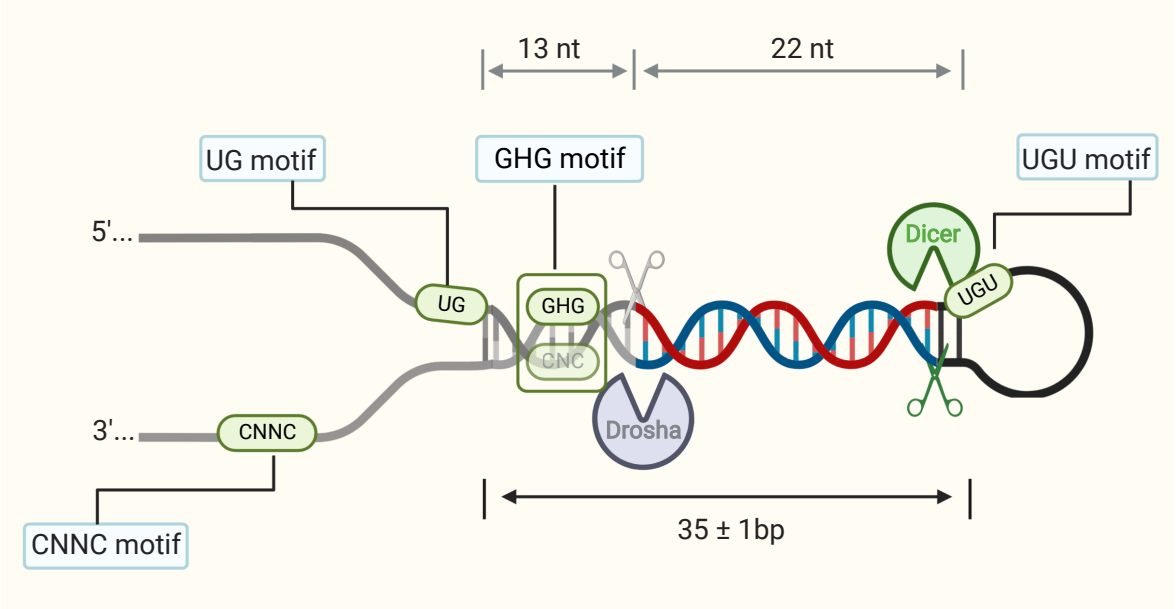



**Figure 7.1:** Pre-miRNA processing. The pre-miRNA stem has an optimal length of 35±1 bp with few mismatches and bulges. There is a basal UG motif, and apical UGU motif, a flanking CNNC motif (N being any nucleotide), and a mismatched GHG motif (H is A, C, or U). Drosha cleaves the pre-miRNA and the hairpin is further processed by Dicer to result in the final mature miRNA of 22 nucleotides. The figure is drawn after Bartel[317].

(nt) long[317;342]. They act as regulatory agents of mRNAs leaving a mark in almost all biological functions of the human body. Not only are they found influential in organ developments, but they are also associated with diseases. The genes lin-4 and let-7 were the first to be identified as part of a class of small RNAs to perhaps have a role in mRNA expression regulation[343;344]. Quite a few of these RNAs are located in intronic regions of protein-coding and non-coding genes. A polycistronic transcription unit is evinced frequently where several miRNA genes are found to be located in close proximity of each other[318]. MYC, ZEB1, ZEB2, MYOD1, and p53 regulate miRNA expression, so does DNA methylation and histone modification[345;346]. Those genes sometimes have several transcription start sites, but can share promoters with protein-coding genes[318;347].

The mature miRNAs are transcribed from stem-loop regions of longer RNA transcripts[317]. These forerunners of miRNAs are known as pri-miRNAs from where miRNAs are transcribed by the enzyme RNA polymerase-II (pol-II), the same enzyme that transcribes mRNAs. It has been noted that some viral miRNAs can be transcribed by RNA pol-III, like miR-142[318]. Drosha endonuclease and the protein DGCR8 constitute the hairpin subtrate Microprocessor, which is formed after a pri-miRNA region folds back on itself[348]. The to-be-folded region generally features an unstructured terminal loop, single-stranded segments at the base of the hairpin, and a 35±1 base pair stem[317;349;350]. There is evidence of four sequence motifs that are potentially present in human pri-miRNAs having a role to play in the processing mechanism. A `UG` motif and a flanking `CNNC` (N stands for any nucleotide) motif in the basal region are present; the splicing factor SRp20 binds to the latter. An apical `UGUG` or `UGU` motif is another one that can be located in the terminal loop. A mismatched GHG (H: A, C, or U) motif has been found to exist in positions relative to Drosha cleavage sites[317;318;351]. The precursor miRNA (pre-miRNA) is formed when Drosha (containing two RNase III domains) cuts

each strand of the pri-miRNA hairpin stem. A 60 nt stem-loop is created which is then exported to the cytoplasm by protein exportin 5 (EXP5) binding to a nuclear protein RAN-GTP[352;353], where it is once more subjected to another endonuclease, Dicer[354]. Dicer creates the miRNA duplex by slicing both the strands near the hairpin loop. One of the strands becomes the guide strand of the silencing complex, denoted as miRNA, and the other, known as miRNA*, is discarded. As a result of the slicing actions of Drosha and Dicer, the duplex retains a 2 nt 3' overhang on each end[317;355]. Figure 7.1 shows pre-miRNA processing.

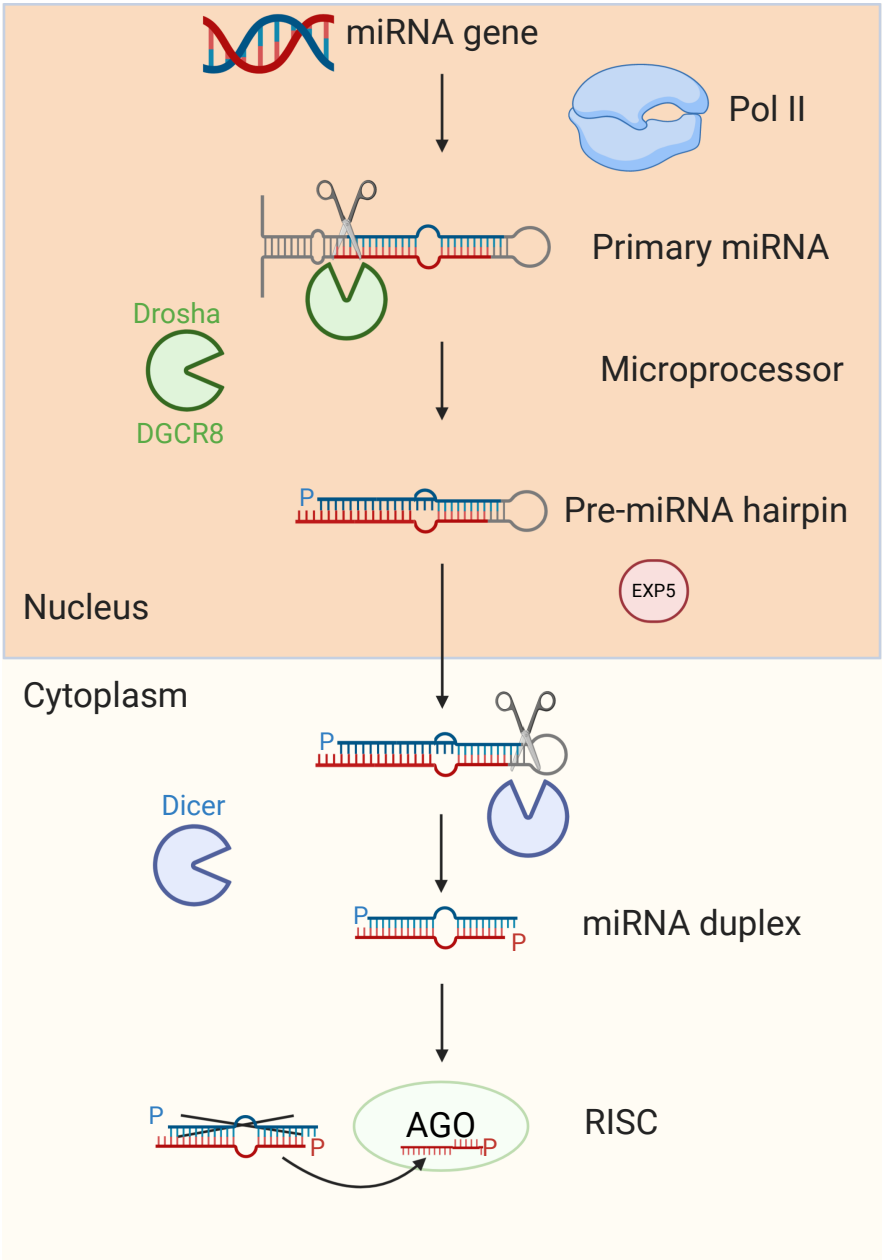



**Figure 7.2:** Canonical transcription pathway of miRNAs. RNA pol-II transcribes the pri-miRNA (originating from exons and introns of non-coding transcripts or pre-mRNAs) and is processed by the microprocessor (containing Drosha) to form the pre-miRNA hairpin, which is then transported to the cytoplasm by Exportin 5 and RAN-GTP, where it is further cleaved by Dicer. The mature miRNA is one strand of the duplex and is loaded onto AGO to form a RNA-induced silencing complex (RISC), while the other strand is degraded. The figure is drawn after Bartel[317].

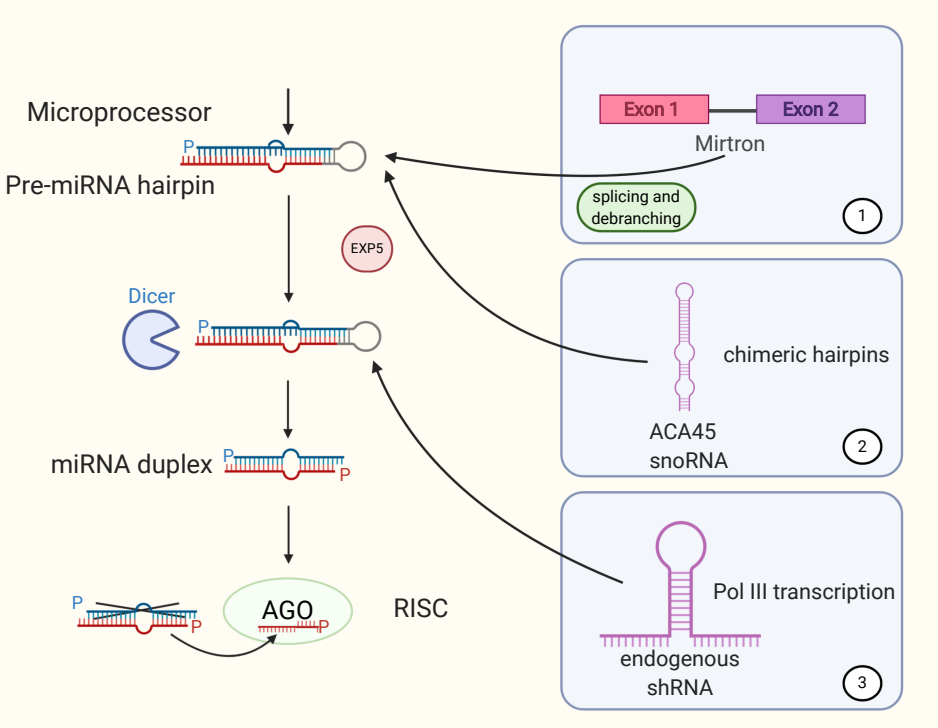



**Figure 7.3:** **(1)** The spliceosome-induced pre-miRNA hairpins called mirtrons that bypass Drosha. **(2)** Chimeric pre-miRNA hairpins also bypass Drosha. **(3)** Endogenous shRNAs are transcribed by RNA pol-III and transported directly to the cytoplasm to be processed by Dicer. This figure is drawn after Bartel[317].

The duplex then binds to an Argonaute (AGO) protein to form the mature miRNA silencing complex, called RISC for RNA-induced silencing complex, thereby unwinding the duplex and releasing the passenger strand, subsequently mostly discarded[356]. The duplex strand which contains a 5'-nucleoside monophosphate in the 5' end that is most favourable to bind to the AGO protein is retained[317;357]. The AGO protein is responsible for escorting factors inducing post-transcriptional and translational repression, mRNA degradation[318;358]. The canonical transcription pathway is shown in Figure 7.2. Some non-canonical miRNA genes produce pre-miRNA hairpins due to the spliceosome instead of Drosha and are called mirtrons (Figure 7.3). These mirtrons are transported to the cytoplasm by Exportin5 as well, however some pre-miRNAs, like 7-methylguanosine (m7G)-capped pre-miRNA, are transported by Exportin1. Endogenous short hairpin RNAs (shRNAs) give rise to some miRNAs that are transcribed by Drosha, but do not undergo Dicer processing due to their length, instead they require AGO2[359]. Certain miRNAs are also transcribed from the same genomic locus as small nucleolar RNA (snoRNA) precursors[317;318;360]. The miRNAs thereafter target other RNAs, mostly mRNAs, to regulate their post-transcriptional activities by pairing to target sites. The seed region of the miRNAs, 2-8 nucleotides at the 3' end, is important for target recognition and bind to the target RNAs to repress them. The nucleotides at position 13-16 are part of the machinery, too, but not as crucial. The diversity of seed regions and target RNAs lead to miRNA genes being clustered into different families[316]. It is estimated that miRNAs target more than 60% protein-coding genes[318;359].

MiRNAs have been detected in various biological fluids such plasma, cerebrospinal fluid, saliva, and breast milk, among others[359;361–363]. It has been reported that

oncogenic miRNAs in breast cancer cells are found in exosomes secreted by IL4-activated macrophages. Also, Docosahexaenoic acid, or DHA, induced exosomes carrying miRNAs to inhibit tumour angiogenesis[364]. miR-105 is involved in metastatic breast cancer cells. Also involved in breast cancer are miR-29a, miR-181a, and miR-652[365]. miR-342-3p and miR-1246 induce metastasis in oral cancer cells. miR-21-3p was found to promote the proliferation and migration of fibroblasts, which in turn led to accelerated wound healing[359;366]. Absence of Dicer in retina can lead to under-expressed miRNAs (the cluster miR-183:miR-96:miR-182) resulting in retinal disorder. The miRNA miR-133b is associated with cerebral ischemia. miR-486 has emerged to a tumour suppressor for NSCLC[367;368].

### 7.2.2 Small nucleolar RNAs

Small nucleolar RNAs (snoRNAs) are a specific group of non-coding RNAs occurring in the nucleolus of a cell. They can be called as a sub-group of small nuclear RNAs which reside in the nucleoplasm and the nucleolus of a cell and are responsible for tRNA and mRNA maturation[369] (Figure 7.4). Some of the functions are conserved in eukaryotes as have been demonstrated through evidence of structural homologues. SnoRNAs are normally found in the intronic regions of protein-coding genes as well as long non-coding genes. As a result, splicing leads to synthesis of the snoRNAs, which in turn get involved in various RNA regulatory mechanisms within the nucleus. These small RNAs are present in abundance in all eukaryotes. Their primary role lies in chemical modification and post-transcriptional processing of ribosomal RNAs[7;370]. mRNAs contain certain structural features that lack in snoRNAs, such as m7G cap at the 5′ end and the polyadenylated 3′ end, which may explain localisation of snoRNAs in the nucleus unlike mRNAs, although both RNAs are transcribed by RNA pol-II[322].

Small nucleolar ribonucleoprotein complexes, or snoRNPs are formed when snoRNAs bind to specific proteins and take part in post-transcriptional modifications. C/D box and H/ACA box are the two main classes of snoRNAs defined with respect to several factors, including sequence motifs, binding partners and secondary structural elements. C/D box snoRNAs are of length 60-200, whereas H/ACA box snoRNA molecules are larger going up to 300 nucleotides[370;372]. The former class of snoRNAs are characterized by the presence of sequence motifs, called C and D boxes, which are highly conserved. The C box refers to a canonical motif `RUGAUGA`, where R is a purine (A or G). The D box denotes a canonical motif `CUGA`. The C and D boxes are present near the 5 ' and 3 ' ends, respectively and form a *kink-turn*, or k-turn, motif inside the folded RNA molecule, a hairpin-hinge-hairpin-tail protein binding secondary structure[370;372], which becomes the binding site for snoRNP proteins. The lowly conserved C ' and D ' boxes can be observed towards the middle of the RNA[7]. The C/D box snoRNAs guide 2 '-O-ribose methylation of certain rRNA residues by forming a long helix and the 7 to 21 guide region lies upstream of the D/D ' boxes[371;373;374] RNA methylation takes place at the $5^{th}$ nucleotide upstream from the D/D ' boxes[371;375;376]. Methyltransferase fibrillarin, 15.5K, Nop56 and Nop58 are the proteins binding to C/D box snoRNAs to form snoRNPs. Fibrillarin primarily controls substrate methylation, while the other ribonucleoproteins are responsible for maturation of the snoRNAs[370;374;377].

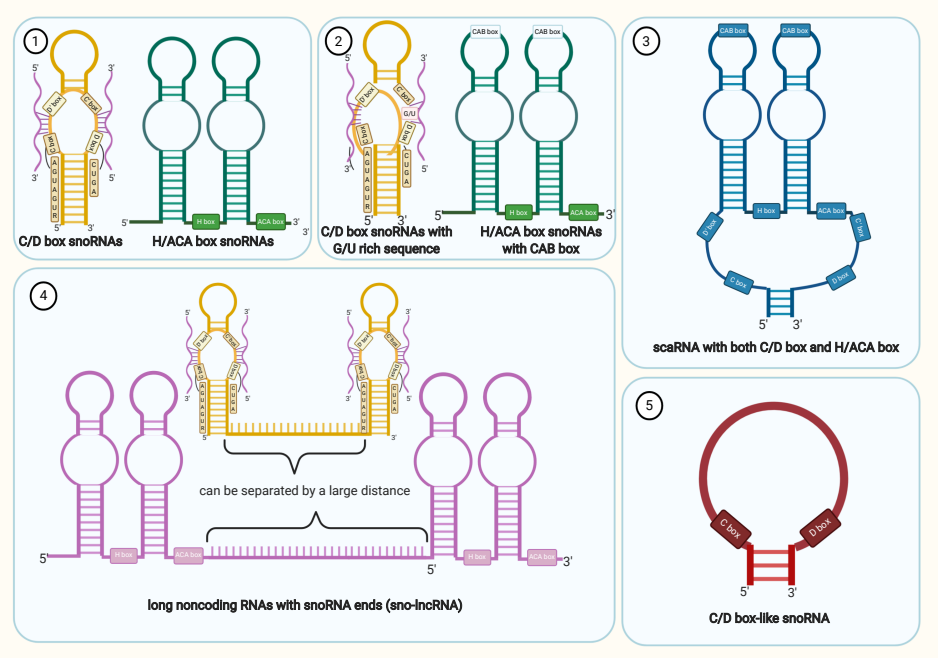



**Figure 7.4:** Different types of snoRNA. **(1)** A C/D box snoRNA with `RUGAUGA` and `CUGA` motifs and a H/ACA box snoRNA. **(2)** Cajal-body associated snoRNA with specific localisation motifs. **(3)** A hybrid snoRNA with both C/D and H/ACA boxes. **(4)** SnoRNA-ended lncRNA. **(5)** Extremely short C/D box-like snoRNA. This figure is drawn after Jorjani et al.[371].

The longer H/ACA box snoRNAs are characterized by the presence of a canonical motif `ANANNA`, where N can be any nucleotide, called a H box and a box containing a trinucleotide pattern, or 3-mer, `ACA` located at the 3' end. This group of snoRNAs have a well-defined structure: two hairpins connected by the H box. The ACA box terminates the second hairpin. The H/ACA box snoRNAs guide pseudouridylation of rRNA residues which are executed through RNA-RNA interactions in the internal stem loops within the two hairpins with the target RNA. Pseudouridine transferase dyskerin[7;374;378], Nhp2, Gar1, Nop10 are the four proteins combining with H/ACA box snoRNAs to form snoRNPs[371]. Dyskerin converts uridines to pseudouridines at 14 to 16 nucleotides upstream of the H and ACA boxes. Two short duplexes are formed by the snoRNAs aligning with the target sequence(s), which are RNA polymerase (pol)-II transcribed spliceosomal RNAs[370;372;379].

Apart from the two major classes of snoRNAs described above, there is a subset of snoRNAs known as small Cajal body-specific RNAs (scaRNAs), which is to be mentioned in passing. They are located in small membrane-less subcompartments of the nucleus known as Cajal bodies and act in post-transcriptional modification of snRNAs. They are normally larger than the two other classes of snoRNAs, contain all the singular/distinctive boxes of those, besides CAB boxes containing a sequence motif `UGAG`. RNA pol-II specific spliceosomal snRNAs are subject to modification in Cajal bodies[7;371].

Certain snoRNAs, like SNORD3, SNORD13, SNORD14 and SNORD22, play roles in rRNA precursors cleavage[370;380;381]. 2'-O-methylation and pseudouridylation of RNA pol-I transcribed rRNAs and RNA pol-II and -III specific spliceosomal snRNAs are areas where the snoRNAs play important roles that have been verified[372;382]. A

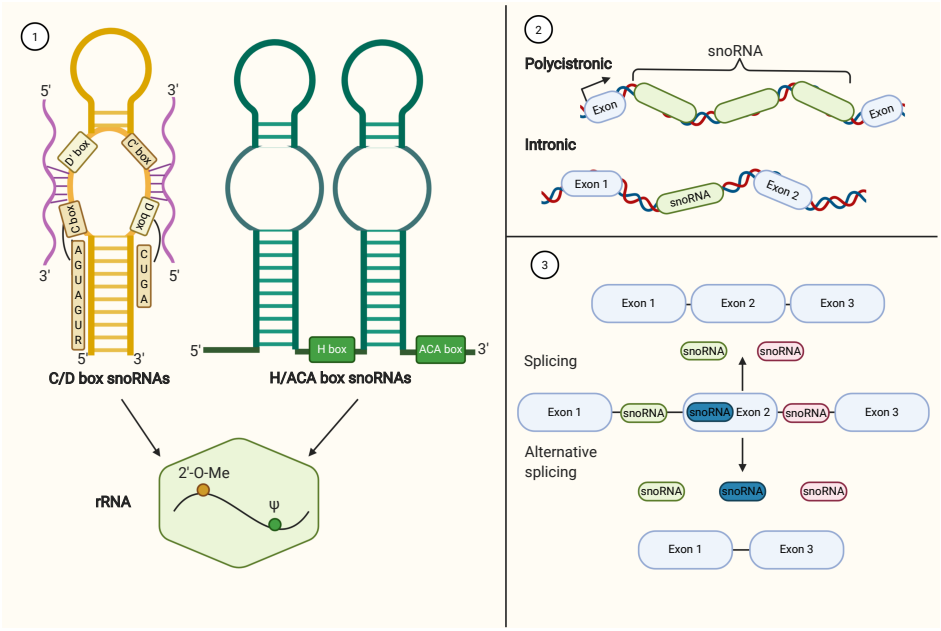



**Figure 7.5:** Functions and origin of snoRNA. **(1)** C/D box snoRNAs and H/ACA box snoRNAs guide 2'-O-methylation and pseudouridylation, respectively. **(2)** Either one or multiple snoRNA genes can be transcribed from introns. **(3)** Alternative spliced leads to snoRNA expression regulation. Only alternative splicing gives way to all three snoRNAs. The figure is drawn after Kufel and Grzechnik[322].

H/ACA box snoRNP is telomerase, which is a RNP reverse transcriptase appending telomeric DNA repeats to chromosomes, that is known to be associated to dyskeratosis congenita, a genetic disorder, through mutations in the H/ACA box.[372] Another snoRNA, SNORD115 (of the C/D box class), regulates alternative splicing of serotonin receptor 5-HT2cR[371;383;384] which is associated with the genetic disorder Prader-Willi Syndrome. It suppresses the efficiency of ADAR2 mediated RNA editing of 5-HT2cR. SNORD115 has also been found to regulate alternative splicing of pre-mRNAs like DPM2 by generating shorter RNAs[385]. The splicing regulator Fox is sequestered by SNORD116[71]. SNORD27 and SNORD88C contribute to regulation of alternative splicing of several pre-mRNAs, with the former being busy with E2F7 transcription factor[370]. SNORD88C contains complementary sequences to a number of pre-mRNAs in its C' box, such as FGFR3, thus regulating alternative splicing[386]. ACA11 downregulates ribosomal protein genes to suppress oxidative stress[387]. Besides rRNA methylation, snoRNAs are increasingly being found in cancer cells and the phenomenon leads researchers to think that they might have oncogenic functions. The gene FBL and snoRNAs have been discovered to be overexpressed in breast and prostate cancers. FBL is overexpressed through the oncogene Myc which led to p53 suppression; however, ribosomal proteins L5 and L11 bind to MDM2 as a direct result of accumulation of p53 due to snoRNA knockdown induced cellular stress, which stabilizes p53[370;388;389]. In non-small cell lung cancer (NSCLC) p53 expression is regulated by SNORA42. Besides SNORA42, SNORD33, SNORD66, and SNORD76 are also overexpressed in NSCLC. SNORNA47, SNORA68, and SNORA78 can be utilised to predict overall survival in NSCLC[370;390]. RUN43, RUN44, and RUN48 are visible in breast cancer and are downregulated[391;392]. Some snoRNAs have been reported to also be effective in predicting prognosis in T-cell lymphoma and B-cell chronic lymphocytic leukemia[370;393]. SnoRNAs transcribed from GAS5 are controlled by p53 associated signalling pathways in colorectal cancer[139;394]. SnR4 and snR45 guide rRNA acetylation and SCARNA97

contribute to tRNA methylation[395].

## 7.3 Functional duality evident in miRNA and snoRNA genes

Certain snoRNAs exhibit functional duality and can be found playing roles in miRNA induced gene silencing. Some precursor snoRNAs play host to miRNA genes; their genomic loci overlap, hence, at times, their features overlap as well. In addition to standard wet lab experimental procedures, with the advent of deep sequencing techniques it is becoming easier to analyse longer transcripts which are processed or degraded to form smaller RNAs. Bioinformatic techniques can predict functions and gene regulatory mechanisms of these smaller RNAs[396]. A few snoRNA genes have been reported to produce smaller molecules with miRNA-like functions. Actually, several small RNAs have been found to be transcribed from snoRNA precursors[7;397;398]. SNORD28 is a p53-repressed snoRNA which can bind to AGO protein and exhibit miRNA-like properties. The other transcript that is transcribed from the gene *SNHG1* is snoRNA-miR-28 which acts as an miRNA. snoRNA-miR-28 binds to and inhibits the expression of TAF9B, which is an mRNA, promoting MDM2 binding to p53, impairing the stability of p53[399]. A miRNA, miR-768-5p, generated from the snoRNA gene SNORD17 binds to YB-1[370]. When miRNA precursors are cleaved from pri-miRNAs by Drosha, they are generally transported to the cytoplasm to undergo further processing by Dicer. However, some mature miRNAs localise in the nucleus, specifically in the nucleolus[400;401]. Both groups of RNAs are predominantly found in introns of longer genes and they have independent transcription units. Exo- and endo-nucleases are necessary for subsets of the two RNA groups along with exosome functionality[7;402]. The human serotonin receptor 2C, HTR2C, has been shown to encode both miRNA and snoRNA genes. The miRNAs miR-448, miR-1264, miR-1298, miR-1911, and miR-1912 are derived from its locus, so is snoRNA HBI-36. The RNase III enzyme Dicer reportedly processes snoRNAs such as scaRNA ACA 45 in human[7].

The Argonaute proteins, AGO1 and AGO2, are essential in the RISC mechanism by forming complexes by binding to miRNA duplexes. Nop56, a C/D box snoRNP core protein has been identified in AGO1, and fibrillarin, the C/D box snoRNP component responsible for substrate methylation has been reportedly found in AGO2[403]. There are plenty of instances of miRNA precursors which contain structural properties of C/D box and H/ACA box snoRNAs, with a few precursor molecules binding to fibrillarin and dyskerin, respective core snoRNP proteins for the two most prominent classes of snoRNAs. Mir-549 gene's precursor transcript has the characteristics of a H/ACA box snoRNA and was predicted to target a rRNA pseudouridylation site. Mir-605 has shown a similar property binding to the core snoRNP protein diskerin, which suggests that it can also be annotated as a H/ACA box snoRNA. It inhibits MDM2 expression while over-expressed and its involvement in the p53 regulatory network are more telling signs of its miRNA characteristics[404]. MiR-140, mir-151, and miR-215 had also been shown to display H/ACA box snoRNA-like attributes in the work by Scott et al.[405]. They had reported the presence of poly(A) tails, target site duplications and the presence of transposable elements in the genomic locations of those miRNAs. Another research

work showed that some pre-miRNA molecules, including miR-28, miR-31, and let-7g, was predicted to have a semblance to structures of known C/D box snoRNAs. The primary structure of the molecules were consistent with characteristic C/D box snoRNA attributes; all were observed to have the C box and the D box and folded as C/D box snoRNAs would. Besides that, all the pre-miRNAs had a tendency to bind to fibrillarin and localised in the nucleolus[7;405]. The C/D box snoRNA U3 functions as a precursor to miR-U3. It includes all the common properties of a C/D box snoRNA; however, it is transported to the cytoplasm where it undergoes processing by Dicer and binds to AGO protein[396]. However, several snoRNA genes have been known to have copied their sequence from one genomic site to another, effectively shifting their allegiance to another host gene in the course of evolution. As a result, different genes in different organisms play host to orthologous snoRNAs[406]. This shows that besides biogenesis and processing pathways, there have been observed similarities in the protein interaction components of the two RNA groups.

## 7.4 Building the classifier

The task at hand was essentially a multi-class classification in machine learning paradigm. There were three classes, namely, SNHG, MIRHG, and NoHG, covering the three different types of sequences, or sub-sequences, that this research project set out to distinguish. Four different collections of sub-sequences were curated into four datasets and a classifier was trained on each dataset with a variety of features. In this section the datasets will be described at the beginning, following which the features will be explained. Details of the machine learning framework will be dealt with next, both supervised and unsupervised methods.

### 7.4.1 Datasets

For a comprehensive investigation into the separability of SNHGs, MIRHGs, and NoHGs, the data was curated at the onset from available annotation datasets. MiRNA sequences were collected from miRBase[407]. The miRBase release 22.1 included annotation for 271 different species, combining 38,589 hairpin precursors and 48,860 mature sequences in one place. It contains 1,917 annotated hairpin precursors, and 2,654 mature miRNAs for human. SnoRNA sequences were retrieved from two datasets: Human snoRNA Atlas[371] and snoDB[408]. SnoRNA Atlas contains 1,118 annotated human snoRNAs, whereas snoDB reports 2,064 snoRNA genes. LncRNA sequences were accrued from GENCODE v.33[5], in which there are 17,952 lncRNA genes listed.

Each of the annotation databases provided coordinates of the genomic location of the sequences in GTF files. The data from miRBase included the exact coordinates of not only precursor miRNAs but also of mature miRNAs. To identify MIRHGs, the simplest step was to extract those coordinates and check if there was any overlap with coordinates of the lncRNA genes. To perform this step of computation the `bedtools suite` was the most suitable tool at hand[409]. The suite offers an array of functionalities to manipulate transcriptomic data using a few intuitive commands.

```
bedtools intersect -a <phastCons.bed> -b <exons.bed>

bedtools merge -i <exons.bed>
```

Two examples of `bedtools` commands used for identifying overlaps between the annotation files of lncRNAs and the smaller RNAs as well as extracting the correct conservation scores at each base position for all the host and non-host genes. The *intersect* command takes two files as inputs and searches for overlaps (defined by the *-a* and *-b* switches). The *merge* command does exactly what it stands for: it merges overlapping features into a single feature.

To extract fields defining lncRNA sequences from GENCODE, the genomic coordinates of the lncRNAs after the merge operation were mapped onto the FASTA files provided on the GENCODE website and the respective sequences were extracted. All snoRNA instances were extracted from the GTF files of the two databases using custom scripts in `Python` programming language.

The `jupyter` environment was utilised in conjunction to be able to have greater control over the program codes and visualise the outputs instantly[410;411]. The packages `pandas`[412;413] and `numpy`[414] within the Python language framework were extensively used for data manipulation and visualisation. The `scikit-learn`[242] was used to perform the supervised and unsupervised machine learning tasks and generate metrics. The `keras` package[415] within the `TensorFlow` framework[236;237] was deployed to design a convoluted neural network (CNN) based autoencoder instead of a classical clustering approach.

For MIRHGs, the payload was defined for both intronic and exonic miRNA precursors with 100nt flanking sequence (Figure 7.6). Since all snoRNAs are intronic, no further processing was needed for SNHGs. To define the exonic part of lncRNAs, the exon/intron annotation provided by GENCODE was used. For training and testing, the datasets were always balanced to match the smallest number of sequences available for any given class, to avoid prediction artefacts. Table 7.1 provides an overview of the number of sequences in every dataset.

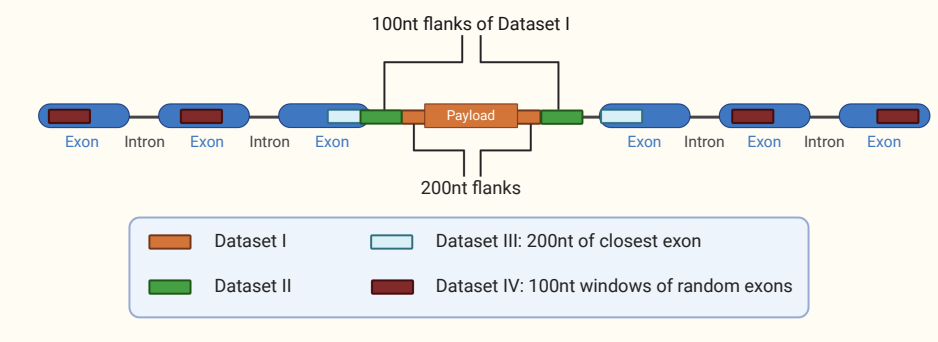



**Figure 7.6:** Datasets creation. A schematic of the datasets curated for this study and their distribution over the gene body of a generic host-lncRNA. Dataset I consists of the payload and 200nt flanking sequence. Dataset II flanks Dataset I by 100nts. Dataset III consists of the first 100nts of the exon closest to the annotated payload. Dataset IV consists of non-overlapping 100nt windows taken from random exons of the host-lncRNA.

### Dataset 1

The first dataset included the payload sequence in the collection of sub-sequences. The smaller RNA lodged in the longer RNA sequence must have all the processing information necessary in the sequence itself was the idea behind creation of this dataset. In case of miRNAs, the processing machinery of Rnase III enzymes, Drosha and Dicer, begins well before the commencement of the actual pre-miRNAs. The entire precursor genes contained within the lncRNA sequences were considered. To accommodate processing information 100 nucleotides (nt) from both ends of the payload sequences was taken additionally, making the mean length of the sub-sequences of MIRHGs and SNHGs to be around 500 nt.
If a payload sequence happened to occur in close proximity of the terminal ends of the lncRNA gene, so that one or both of the flanking regions was shorter than 100 nt, the payload sequence was discarded altogether to give equal weight to up and downstream regions. After selection, the number of SNHGs and MIRHGs amounted to 345 and 400, respectively. The negative control set contained 400 randomly selected 500 nt sub-sequences of lncRNAs, which did not have any overlaps with either SNHGs or MIRHGs.

### Dataset 2

To investigate the influence of the information on the classifier derived directly from the payload, the next dataset was curated with sequences excluding the regions defined in Dataset 1. This was necessary to observe, if the classifier could detect crucial information outside the immediate vicinity of the precursor genes. Dataset 1 provided the classifier with the bias of actual payload sequence; it would glean information from the precursor genes. To explore the possibility of detecting the classes of host genes without that bit of information, this dataset was designed. 100 nt windows flanking the regions defined in Dataset 1 were considered. Both snoRNA and miRNA precursors can cluster too closely in the genomic locus, thereby creating a polycistronic unit. Aware of this possibility, overlaps between the extracted sub-sequences were discarded. It must be mentioned here that the focus is not the genic region of the flanks extracted. This also made certain that any potential processing information in close proimity of the genomic context was preserved.
The number of sub-sequences labeled SNHGs and MIRHGs amounted to 690 and 800, respectively. The negative control set was once again a collection of random 100 nt regions of non-overlapping lncRNAs. 750 NoHGs were selected.

### Dataset 3

Since miRNA precursors can reside in both exonic and intronic sequences of the longer RNAs (both protein-coding and non-coding), considering information stored about their processing in the adjacent genic segment is of particular interest. This is also true for snoRNAs; even though they are found in introns, the information encoded in the nearest exons could also be revealing. Dataset 3 was built on this principle. Here, data from

the exons closest to the payload sequences was taken into consideration. Again, 100 nt segments of both upstream and downstream sequences were extracted, in a similar vein as Dataset 2. This implies that all the intronic data surrounding the payload sequence was overlooked and only the exonic data was concentrated upon. In the end, 1,287 SNHGs and 464 MIRHGs were selected. 2,101 exonic regions were also picked up from random lncRNAs, which did not have any overlaps with either of snoRNA or miRNA host gene sequences.

#### Dataset 4

So far, either only the information with the precursor sequences situated within longer lncRNA genes or information from the immediate genomic location was considered . To avoid any localised effect associated with the payload sequence, Dataset 4 was designed. The only similarity of this dataset with Dataset 3 lies in the choice of exonic regions. However, in this case, the exonic regions selected are not in the close vicinity, i.e., at least more than 100 nts away. For every host gene, multiple, non-overlapping regions of length 100 nt was extracted from the exonic regions only. If the number of segments that could be extracted was less than four, that particular gene was discarded. This was done to have a uniform distribution of sequence information. Both upstream and downstream locations were considered. Since the distance selected was great to avoid any interference with the immediate processing signals, Dataset 4 shrunk. It contained 162 and 168 sub-sequences extracted from snoRNA and miRNA precursor hosting genes, respectively. The same procedure was followed to create the negative control, where 100 nt windows of random exons of random lncRNA genes were selected that did not have any overlap with the positive set. There were 750 sub-sequences labelled as NoHG.

**Table 7.1:** Distribution of sequences in every dataset. Dataset 1 refers to the payload and its flanking sequence. Dataset 2 refers to just the flanking sequences. Dataset 3 and 4 contain exonic information in the immediate neighbourhood and not in close vicinity, respectively. The varying number of sequences depended upon the availability of transcripts subjected to the stringent conditions devised, non-host lncRNAs sometimes outstripping the other classes due to their abundance. For training and testing all datasets were down-sampled to the lowest number of available host gene sequences.

| | **Dataset 1** | **Dataset 2** | **Dataset 3** | **Dataset 4** |
|---|---|---|---|---|
| SNHG | 345 | 690 | 1287 | 162 |
| MIRHG | 400 | 800 | 464 | 168 |
| NoHG | 400 | 750 | 2101 | 750 |

### 7.4.2 Feature engineering

A machine learning algorithm requires certain features to be able to train a model that would help extract meaningful information from the data it is trained upon. Through training, the model becomes more knowledgeable, which is to say it *learns*, and can be used to make predictions on new data. The model is subjected to unseen data during the process of training, which enables it to learn and make predictions, and check the accuracy of the predictions. This learning process is facilitated by features, which, especially in a classification task, enables the model to extract enough common patterns

to distinctly identify the classes. To this end, four categorical features were defined, besides the **sequence** itself, inspired by previous work done in the community, that mostly covered all the information that can be gleaned from a sequence. The defined features were:

1. **$k$-mer counts**
2. **Fickett score**
3. **Sequence conservation scores**
4. **Secondary structure**

Concurrently, the amount of G and C nucleotides present in each sequence was added under **GC content**. Since at least a fraction of the annotated lncRNAs is annotated from incomplete transcript models[66], parameters such as the number of exons, the transcript length, or polyadenylation were not considered in any of the classification tasks. Here, the four features used will be described in detail.

### $k$-mer counts

A $k$-mer is collection of $k$ adjacent nucleotides in a sequence. For example, in a sequence `(...AAGGATCTACCTTGGA...)`, some $k$-mers starting at position 1 can be extracted as follows:

| | |
|---|---|
| 2-mer = AA | (...**AA**GGATCTACCTTGGA...) |
| 3-mer = AAG | (...**AAG**GATCTACCTTGGA...) |
| 4-mer = AAGG | (...**AAGG**ATCTACCTTGGA...) |
| 5-mer = AAGGA | (...**AAGGA**TCTACCTTGGA...) |

It is important to note $k$-mers with smaller $k$s are abundant and may not encode enough underlying genomic information, but $k$-mers with higher $k$ values are rarer and due to their low abundance, they possess more details. Analogously, higher $k$-mers are computationally expensive to calculate.

Inspired by the $k$-mer profile proposed by Kirk et al.[288], a scheme was designed to methodically extract all the information $k$-mers of the sequences could provide. To start with, all the $k$-mers for $2 \leq k \leq 7$ for all the individual sequences were generated. The tool used for this purpose was `jellyfish`[416]. It is a command-line tool which crawls FASTA files containing sequences using a multithreaded hash table. It populates an array of $(key, value)$ pairs that essentially stores each $k$-mer and its frequency for a sequence. Thereafter, it produces an output file for every transcript which, in this case, was read by a custom python script to extract the $k$-mers and their frequencies.

```
jellyfish-linux count -m %d -s 100M -t 10 -C %s -o <kmerfile>.jf

jellyfish-linux dump -c <kmerfile>.jf > <kmerfile>.count
```

These two commands enabled obtaining the $k$-mer frequencies. The *-m* switch asks for the value of $k$ to be counted, using a hash of 100 million elements defined by the switch *-s*, with 10 threads (switch *-t*). The output file (*-o*) was "dumped" using the second command to obtain a human readable list of $k$-mers.

Then, all possible $k$-mer combinations were generated with $k$ values ranging from $2 \leq k \leq 7$. The number of possible $k$-mers can be simply calculated to be $4^k$, amounting to:

$$n(\text{2-mers}) = 16$$
$$n(\text{3-mers}) = 64$$
$$n(\text{4-mers}) = 256$$
$$n(\text{5-mers}) = 1024$$
$$n(\text{6-mers}) = 4096$$
$$n(\text{7-mers}) = 16384$$

Once the $k$-mers were generated, three snapshots of the datasets were created with additional attributes. `key` simply notified the classifier if a transcript contained a particular $k$-mer. It could take the value either 0 or 1, indicating absence and presence of the $k$-mer, respectively. `freq` showed the frequency of a particular $k$-mer of a transcript. `kmer_norm` was the normalised count of the $k$-mer for a particular transcript. It was based upon:

$$kmer\_norm = \frac{freq}{\Sigma_{i=1}^{n} freq_i} \tag{7.1}$$

where $freq_i$ denotes the frequency of that particular $k$-mer across all the transcripts. The three attributes calculated were then combined into one tuple for each sequence:

```
(key, kmer_norm, freq)
```

This tuple was then converted into features using a `MultiLabelBinarizer` class from the `scikit-learn` package. This tuple constituted the primary feature for training the model. Through implementation of the $k$-mer characteristics and normalising the $k$-mer frequencies across all the transcripts, a lot of ground was covered with respect to sequence data.

### Fickett score

As mentioned in the previous chapter, the Fickett TESTCODE is a measure for coding potential of sequences[280]. Having been put to successful application in several tools distinguishing non-coding genes from protein-coding genes, this was implemented in this

work as well, although exclusively non-coding genes were being dealt with. LncRNAs show very little coding potential, however, since the genes considered here encodes smaller RNA genes which participate in gene regulatory mechanisms, too, it was included as a feature set to test if any discernible signal could be found that would help distinguish the three classes. A custom python script was implemented inspired by the work of Wang et al. used in the tool CPAT[281]. The Fickett score was calculated for every instance in each dataset.

#### Sequence conservation scores

The UCSC Table Browser functions as a one-stop shop for accessing and processing a large array of genomic data[417]. Conservation scores for all the sequences in the datasets were obtained from the Browser. The *phastCons* score for a nucleotide is a prediction based upon a phylogenetic Hidden Markov Model (HMM) in its most conserved state[303]. The conservation scores are available in data files containing chromosome coordinates and the scores of the nucleotides in the genomic assembly in a step-based interval. For Datasets 1 and 2, the extraction of the conservations scores was fairly straightforward based upon the genomic coordinates mentioned in the data files and the annotation files at hand. However, for Datasets 3 and 4, scores for the individual exons had to be extracted. The conservation scores was to be utilised in two different ways. Certain published lncRNA detection tools have used the mean conservation score for a sequence as a feature[301;304]. This was one of the approaches taken. Additionally, the scores for a sequence was divided into 5% bins and normalised, which resulted in 20 additional features.

#### Secondary structure

RNA secondary structure is often better preserved than the sequence. Therefore, structural attributes of the sequence was also taken into consideration. Using `RNAplfold`[299] of the `ViennaRNA` package pairing probabilities of the nucleotides were calculated. Three different windows of lengths 60, 80, and 120 were considered for sequence scanning. Furthermore, the windows were also divided into 20 bins, with the nucleotide positions falling into a bin binary encoded. A position was encoded as 1 if it belonged to a bin and the rest of positions were encoded as 0. This approach generated 60 additional features.

### 7.4.3 Feature combinations

The next phase was training a model on the features to accomplish the task of distinguishing the three different classes of lncRNAs in the datasets. However, the training process was split three-way for every dataset. After defining and computing the features, they were combined into three separate groups. The motivation behind this manoeuvre was to observe the efficacy of the features, combinations or lack thereof, by investigating the performance metrics. The three feature groups were as follows:

i Feature Set 1: The first feature group contained only the features derived from *k*-**mer counts**.

ii Feature Set 2: The second feature group included features derived from pairing probabilities of nucleotides, i.e., **secondary structure information**, in addition to the *k*-mer counts.

iii Feature Set 3: The third feature group included all the features from the previous feature groups and the **conservation scores** of the nucleotides in the sequences.

All the three feature sets were further attached with the **Fickett score**, which acted in a binary fashion. The model was trained including and excluding the Fickett score, applicable to all three groups. The actual **sequence** and **GC content** was constantly present for all training runs.

## 7.5 Supervised machine learning

The task at hand is a multi-class classification problem, as there are three different classes a particular sequence (or sub-sequence) could belong to. The sequence can either be a snoRNA hosting lncRNA gene, an miRNA hosting lncRNA gene, or a random lncRNA gene not having any overlaps with the smaller RNA genes. Most lncRNA detection tools have put support vector machines (SVM) to good use, however, it was a binary classification job in most of the cases - separating lncRNAs from mRNAs. Since there were three classes in the present research problem, a random forest classifier was chosen to be trained within the realm of supervised training. The classifier would be trained on the features designed and would learn new patterns in the data for every class, as it would have access to the labels. Random forests are actually an extension of decision trees. Decision tree-based classifiers are normally fast. They traditionally employ a central axis projection technique by constructing a hyperplane dividing the lines that connect two data clusters and identify classes at each decision node. However, the decision tree is approach is prone to overfitting. Random forest-based classifiers grow multiple decision trees by splitting the feature space into random feature sub-spaces and train the individual trees practically on different subsets of the training data. The final combines the accuracy measures from all the trees, thereby avoiding overfitting, i.e., maintaining generalisation accuracy[230;231]. Training a random forest-based model proved useful with regards to the speed of classification as well as obtaining feature ranks. The latter property was used to investigate the most useful features that the model used to reach its decision.

The classification task was designed using the `RandomForestClassifier` class from the `scikit-learn` package. The default number of trees is 100 and it was unchanged for the initial classification. The dataset for the current classification task was split into two parts: 80% of the samples constituted the training set and 20% formed the test set. The training and the test sets were further split into two arrays apiece: `X_train` and `y_train`, `X_test` and `y_test`. `X_train` and `X_test` contained the encoded features for the training set and the test set, respectively. As all the features were numeric by type, they were passed through the `StandardScaler` class, which standardizes the data

into a normal distribution by removing the mean scaling down to unit variance. The standard score of a sample x is calculated as:

$$z = \frac{x - u}{s},$$

where $u$ is the mean of the training samples and $s$ is the standard deviation of the training samples, as described in the `scikit-learn` user manual. Missing values are treated accordingly by the class. The categorical feature of sequence was encoded by the `OrdinalEncoder` class, which converts all discrete features into a string of integers and returns a single column. The `y_train` and `y_test` arrays contained the labels for the training and the test set, respectively. They were also encoded by the `OrdinalEncoder` class.

```
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(n_estimators=100, n_jobs=-1,
random_state=42)
```

For further training, all the available hyperparameters were tuned and optimised for the best results. To divide a training set of size $N$ into $S$ different subsets, so that each $S_i$ contains samples from the original training set and the subsets are so populated that there might be duplicate samples besides a few unique ones in each subset, is known as *bootstrap aggregating* or bagging. Following this principle, the model is trained on multiple new training sets and the outcome is aggregated before performing a plurality vote to choose a class[232]. Accuracy can be improved through bagging. Bagging is an integral part of random forests. For this task, bagging was always enabled. *Out-of-bag error* or oob, was naturally included in the validation step, which enables the bootstrap aggregator to estimate the prediction error of instances not included in the trees used for training to evaluate and improve classification in the next phase of learning. Another tune-able hyperparameter is the *criterion* based upon which a decision tree is split. The *gini* impurity index offers the probability of mis-classification of a particular sample by randomly choosing a label from a node. The *entropy* index calculates the information gain based on a particular node. If all of the samples in a particular node belongs to the same class, the entropy would be zero. Gini impurity is calculated by:

$$G = 1 \ - \sum_{j=1}^{c} p_j^2,$$

where $p_j$ is the proportion of samples belonging to class $c$ for a particular node.
Entropy is calculated by:

$$H = 1 \ - \sum_{j=1}^{c} p_j log_2(p_j),$$

where $p_j$ is the proportion of samples belonging to class $c$ for a particular node and $p \neq 0$.
The number of trees, or *estimators*, were also fine-tuned. The validation was performed on tree sizes [100, 300, 500, 1000].

Feeding all the hyperparameters into the `GridSearchCV` class 10-fold and 5-fold cross-validation of the model was performed. The cross-validation technique splits the

training set into $x$-folds, where $x-1$ sets are used by the random forest classifier to train the model. The remaining split set is then utilised as the validation set for the model to evaluate its aggregated outcome, its predictive power. This process is repeated $x$ times over, essentially training $x$ different models, to optimise the parameters such that at least one of the models fit the training data as well as possible. The `GridSearch` class enables the user to perform this task, where a parameter grid of the hyperparameters can be fed into and the `GridSearch` class instructs the model to cycle through all the available options. For this problem *scoring* metric was set to *accuracy*. Finally, the class provides the best performance based upon the *scoring* metric achieved during the cross-validation tests.

```
param_grid = {'oob_score': [True, False], 'bootstrap':[True],
              'criterion': ["gini", "entropy"],
              'n_estimators':[100,300,500,1000]}

grid_search_def10 = GridSearchCV(
RandomForestClassifier(random_state=42), param_grid,
    scoring="accuracy", n_jobs=-1, verbose=1, cv=10)

grid_search_def5 = GridSearchCV(
RandomForestClassifier(random_state=42), param_grid,
    scoring="accuracy", n_jobs=-1, verbose=1, cv=5)
```

As it happened, for all the datasets, only minor differences in prediction accuracy rates were observed in the given parameter space. The moderate size of the available datasets is probably the reason behind it, since any machine learning algorithm feeds on data, and the more data provided, the better patterns a model can extract from the underlying features.

## 7.6 Unsupervised machine learning

As an alternative approach unsupervised learning was tested on the same four datasets. In unsupervised approach the model does not have access to the labels of the samples and it attempts to decipher new patterns from the instances and their features in the training set. The model essentially makes predictions based upon its own decisions trained on just the features. The first method that was implemented was $k$-means clustering[235]. $k$-means clustering creates $k$ clusters and arranges the different observations into those clusters. Every cluster contains a cluster centroid and the algorithm attempts to tie each observation to a cluster in a way that its distance is at minimum to a centroid. Ultimately, it aims to cluster points into $k$ groups of equal variance. In these project, an attempt was made to distinguish the sequences into three different clusters. `scikit-learn`'s `KMeans` class was implemented to perform this task.

```
KMeans (n_clusters=3, init="k-means++",
        random_state=42).fit(X_train)

km.predict(X_test)
```

A scoring metric was used to evaluate the performance of the clustering algorithm. Besides that, the various attributes provided by the KMeans class was also interpreted for evaluation.

To obtain a visual overview of the data principal component analysis (PCA) was used[233]. PCA generally converts high dimensional data to a lower dimension using singluar value decomposition. It does not scale the input rather centres the data for each feature to the number of dimensions where it perceives most variance. The `PCA` class provides the necessary parameters to carry out an analysis and evaluate them using the in-class attributes.

```
p = PCA(n_components=2, random_state=42)

X = p.fit_transform(X_train)
```

The third approach attempted is based on deep learning. Deep learning works best on enormous amounts of data, however, it was thought to be worth a try to see, if a neural network would recover any discernible patterns from the modest amount of data that could be provided to it. A convolutional neural network (CNN) functions using convolutional layers which contain parameters that automatically extract useful patterns from input data. The convolution kernel processes the feature map, which is the input data, to create a transformed feature map. With multiple layers in one network, a CNN automatically learns features in different stages, following which all the layers collapse into one final outcome layer that contains all the necessary information and can be used for object detection. The convolutional layers filter inputs by utilising the information gained through parameters that are intuitively tuned. A CNN exploits compositional hierarchies to assemble higher level features from lower level features. A convolutional layer detects local summaries of features from the previous layer. A CNN also implements pooling layers for limited translation and rotation invariance, *i. e.* lets little variance in the representations when the previous layer elements vary substantially; by extension, pooling layers also allow more convolutional layers by reducing memory usage. In the end, the network finds the best features suitable for the task. CNNs have been used for signal processing and video processing, but have been found to be most successful in image recognition[202;206]. To this end, the `keras` package within the `TensorFlow` framework was used. The lines of code below give an example of the definition of one convolutional layer and one pooling layer.

```
conv1 = layers.Conv2D(filters=64, kernel_size=(3,3), strides=(1,1),
activation='relu', padding='same')(inp)#input layer

conv1 = layers.MaxPooling2D(2,2)(conv1)
```

With that, the training and testing framework for all the datasets. There were some interesting results that emerged upon analysis of the performance metrics. We will have a look at the insights gained in the next chapter.

# 8

# Can the lncRNA classes be separated?

As explored in the previous chapter, the research question was to find out, if snoRNA host genes (SNHGs) and miRNA host genes (MIRHGs) are distinct classes of lncRNAs beyond their payloads - the smaller RNA genes the lncRNAs are hosting. To discover a similarity in their originating genomic loci would throw some light on the inherent functions of lncRNAs, since snoRNAs and miRNAs are more ancient and more evolutionary conserved and their functions have been better studied. In this research work, the extent of the distinguishing patterns between the host genes and non-host genes (NoHGs) was explored emplying machine learning technqiues. The availability of tools and techniques can quite easily distinguish snoRNAs and miRNAs from each other[339;418;419]. Bearing that in mind, the investigation focussed on whether the distinguishing properties, *i. e.* the distinction of the payloads between themselves and from the non-host genes, still held true, if the payload related information was withheld or hidden from the classification machinery. To this end, a random forest classifier was deployed and trained on all the datasets defined in section 7.4.1 of Chapter 7 and the performance metrics were analysed to determine up to what extent the three classes SNHGs, MIRHGs, and NoHGs were classifiable. The achievable classification accuracy is indicative of the coherence of the RNA classes. As already mentioned previously in section 7.4.3, three different feature sets were designed to train the classifier on.The classification accuracy levels were compared with test results based upon sequence-only features, in particular weighted $k$-mers, and upon feature sets extended by predicted secondary structures and sequence conservation parameters, respectively. The Fickett score was added as an extra feature to each comparison to probe the influence of the coding potentials of each sequence. The robustness of all results is ensured by 10-fold cross-validation (CV) on all data and feature sets, as can be seen in Table 8.1.

## 8.1 Results of classification

The performance of the random forest classifier was analysed using accuracy rates reported by the classifier on a predicted outcome. Furthermore, cross-validation was also employed to determine the robustness of the classifier. In cross-validation, the dataset is split into 10 random parts (for a 10-fold CV) and the classifier is trained on 9 parts, or folds. The last excluded fold acts as the validation set for the model to be tested upon. This whole process is repeated 10 times, that means 10 different models are trained on random parts of the training set and optimised to fit the training data. To inspect the accuracy rate using default parameters, the following piece of code was employed on every dataset for each feature set.

```
rf = RandomForestClassifier(n_estimators=100, n_jobs=-1,
     random_state=42)
rf.fit(X_train, y_train)
y_pred = rnd_clf.predict(X_test)
accuracy_score(y_test, y_pred)
```

`n_estimators` set the number of trees to grow for the random forest.

To inspect the cross-validation accuracy rates, the following lines of code was used.

```
grid_search_def10 = GridSearchCV(
    RandomForestClassifier(random_state=42), param_grid,
    scoring="accuracy", n_jobs=-1, verbose=1, cv=10)

grid_search_def5 = GridSearchCV(
    RandomForestClassifier(random_state=42), param_grid,
    scoring="accuracy", n_jobs=-1, verbose=1, cv=5)

grid_search_def10.fit(X_train, y_train)
grid_search_def5.fit(X_train, y_train)

grid_search_def10.best_score_
grid_search_def5.best_score_
```

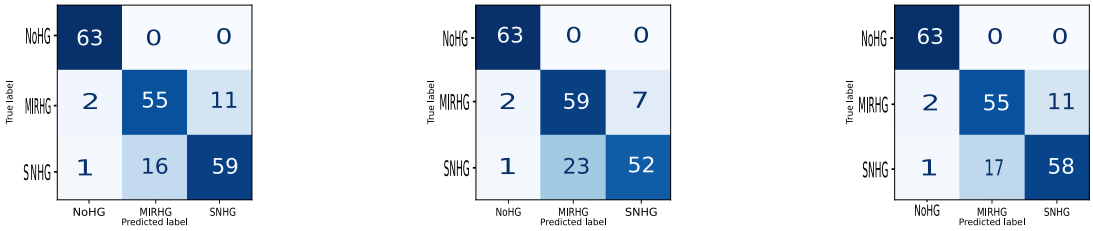
The various hyperparameters to be tuned were made available to the grid search machinery through `param_grid` (Appendix Table A.4). A 5-fold CV was also conducted, however, those results were discarded from the final analysis, since 10-fold CV offers a higher variance of the data with less bias. The cross-validation score reported was based upon the mean cross validated score of the best model performance on the validation set.

The first four subsections will describe the results obtained from the supervised classification task. The prediction accuracy values are compiled in Table 8.1 for reference. The models were retrained several times to minimise the effect of the stochasticity of the classification and evaluation process. The values presented in the table are the most robust results. Confusion matrices were generated for each approach based upon the prediction performance of the model on the test set and are an excellent way to visualise the separation between the classes.

### 8.1.1 Classification on sequences including payload

The intention behind this strategy was to feed the classifier sequences with payload included and expect that the classifier learned enough to distinguish those from sequences without any hosted genes. A subsequent motivation was to have the classifier recognise the patterns of two distinct payloads and identify them. Hence, this approach quite understandably was to become the yardstick of the feature selection strategies. Dataset 1 served this purpose, including sequences that contained the smaller RNA genes as well as both upstream and downstream flanking segments. 345 samples of each class were selected to achieve a balanced dataset (Figure 8.1).

With default parameters, using only sequence derived features, the accuracy rate of the model was a shade above 83%. However, on computing 10-fold CV (Table 8.1), the accuracy value achieved was above around 85.5%, which showed that random forests were indeed suitable for the classification task at hand.



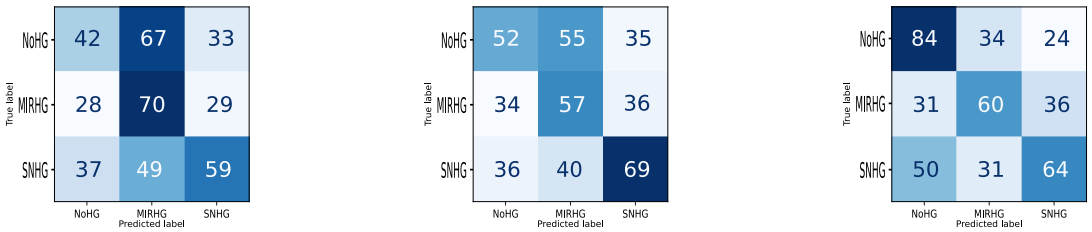

**(a)** Feature set 1 **(b)** Feature set 2 **(c)** Feature set 3

**Figure 8.1:** Confusion matrices for each of the three feature sets for dataset 1. The y-axis depicts the true values and the x-axis the predicted values. The diagonal depicts the true positives. ∘ Legend guide: [lnc: NoHGs • mir: MIRHG • sno: SNHG]

An interesting finding, however, was the decrease in accuracy levels when features derived from secondary structure and conservation scores were also considered. This was thought to be probably due to the high structured-ness of both snoRNA and miRNA payloads and their similar conservation patterns. It is imperative to also take into account that only features derived from the sequences, *i. e.* originating from $k$-mers, constituted the most important features according to ranking of features (Appendix Table A.3). Finally, it was noted that a perfect separation of snoRNA and miRNA precursors genes could not be achieved or expected due to the presence of a group of ncRNAs that appear to be in transition between those two groups[7;405;420]. Further performance metrics are available in the Appendix (Table A.5).

### 8.1.2 Classification on flanking sequences

Since pri-miRNAs cover quite a bit of the genomic loci before being cleaved to pave way to precursor miRNAs (which are then further processed by Drosha to reach close to the length of a mature miRNA), it can be naturally assumed that the flanking sequences of the payload will contain essential information about the hosted miRNA




**(a)** Feature set 1 **(b)** Feature set 2 **(c)** Feature set 3

**Figure 8.2:** Confusion matrices for each of the three feature sets for dataset 2. The y-axis depicts the true values and the x-axis the predicted values. The diagonal depicts the true positives. ○ Legend guide: [lnc: NoHGs • mir: MIRHG • sno: SNHG]

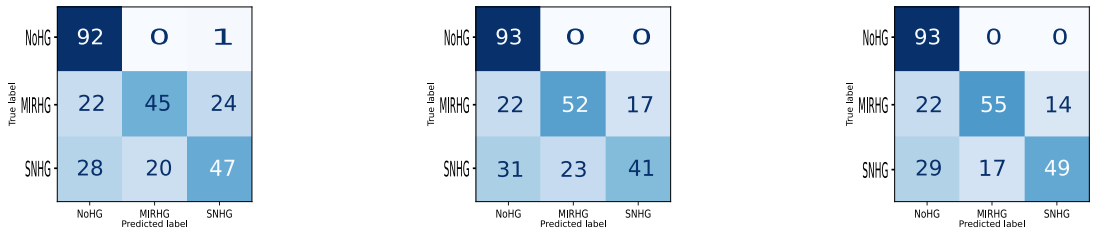
in the MIRHG. Similarly, SNHGs will also encode information around their hosted snoRNAs about its processing. As in the case of Dataset 1, the classifier was made aware of that information and it is expected that it picked up on those signals. Dataset 2 consisted of sequences to solely inspect, if it is necessary to have the payloads available to the classifier, or is it enough to have the sequences from the immediate vicinity from whom the classifier would pick up the processing related signals in order to distinguish between the hosted genes. Dataset 2 contained sequences of length 200 nt including 100 nt of flanking sequences from either side of the payload without any overlaps. Random 200 nt sub-sequences of NoHGs were selected as the negative set. 690 samples of each class were selected to have a balanced dataset to train the classifier on (Figure 8.2).

It was observed that exclusion of the actual payload sequences had a strong impact on prediction accuracy. The classifier trained on the default feature set with default parameters returned just over 39% accuracy value. The performance evaluation through 10-fold CV also confirmed the lack of existing patterns in the data as the accuracy levels plummeted to less than 45% in contrast to the model performance for Dataset 1 (Table 8.1). Further performance metrics are available in the Appendix (Table A.6).

For comparison, a uniform random sampling would have achieved accuracy values of 33%, since a balanced three-way classification problem was considered. However, integration of secondary structure and conservation features indeed achieved a moderate increase in accuracy values to just above 50%. Hence, it can be concluded that the sequences in the vicinity of the actual payload are insufficient to reliably identify the payload type. Ranking of features according to their importance in the classification task showed that upon exclusion of the payload, sequence and sequence conservation become important features for the task, instead of just information derived from $k$-mers (Appendix Table A.3).

### 8.1.3 Classification on exons adjacent to the payload

After handling two use cases with the payload included and the genomic region in the immediate vicinity of the payload, it was time for looking at annotated information around the payload. Dataset 2 included information mostly from the intronic regions of the lncRNAs, since all human snoRNAs and most miRNAs are transcribed from




**(a)** Feature set 1 **(b)** Feature set 2 **(c)** Feature set 3

**Figure 8.3:** Confusion matrices for each of the three feature sets for dataset 3. The y-axis depicts the true values and the x-axis the predicted values. The diagonal depicts the true positives. ○ Legend guide: [lnc: NoHGs • mir: MIRHG • sno: SNHG]

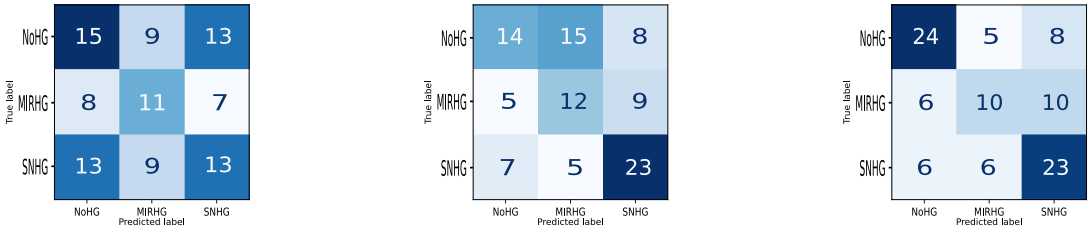
intronic regions of protein-coding or other longer non-coding genes. However, functions of the mature host genes or other lncRNAs are presumably encoded in their exonic sequences. Furthermore, sequence and structure features involved in splicing as well as the initial processing of exonic miRNAs are likely to be found in the mature lncRNA transcript. Concordantly, Dataset 3 was designed to focus exactly on those cases. It comprised of the 200 nt segment of the exonic sequence flanking the payload-bearing intron in case of an intronic payload or the miRNA precursor in case of exonic miRNAs. The down-sampled dataset contained 464 samples of each class.

In comparison to Dataset 2, the accuracy rate increased in this case (Figure 8.3). Evaluating the model through 10-fold CV reported close to 67% accuracy rate, whereas 62% was achieved with the default parameter set (Table 8.1). With the addition of secondary structure and conservation scores, the accuracy level climbed up to a shade more than 70%, which could be interpreted as the presence of more reliable processing information in the sequence fragments selected that what was detected by the classifier in the immediate flanking segments of the payload. Further performance metrics are available in the Appendix (Table A.7).

For miRNAs this can be explained in particular by sequence motifs associated with microprocessor activity, for instance, the flanking `CNNC` motif, the `UG` and `GHG` motifs at the base of the hairpin at Drosha cleavage sites[317;421;422] (Fig. 7.1). In case of snoRNAs, this maybe explained by their connection to splicing. Some snoRNA host genes are suggested to be spliced so that their intronic snoRNA genes are released and transcribed[322;423].

### 8.1.4 Classification on random exonic sequences

Assessing the results from the analyses of the previous datasets, it can be inferred that the exonic regions flanking the payloads contain some amount of information. However, it was not clear whether the regions contained signals to distinguish the three classes based on the payload, or they simply contained signals for a global distinction between the three classes of lncRNAs. Dataset 4 stemmed out of this idea and it constituted of random fragments of exonic regions excluding the payloads and their immediate neighbourhoods. Since these regions are much less likely to contain sequence signals




**(a)** Feature set 1 **(b)** Feature set 2 **(c)** Feature set 3

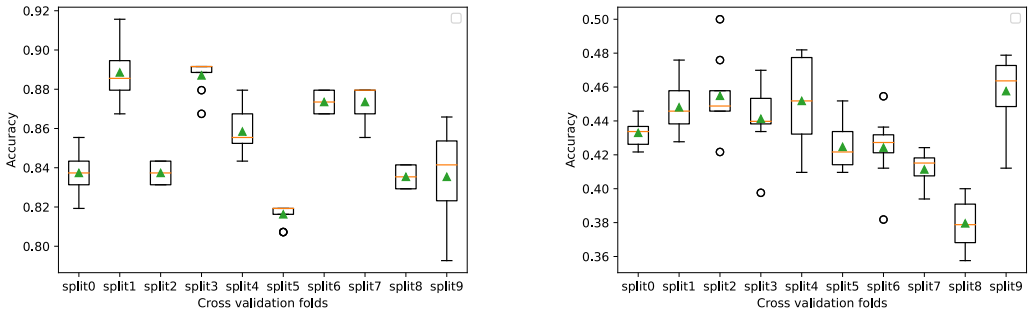
**Figure 8.4:** Confusion matrices for Dataset 4. The y-axis depicts the true values and the x-axis the predicted values. The diagonal depicts the true positives. ∘ Legend guide: [lnc: NoHGs • mir: MIRHG • sno: SNHG]

directly related to the processing of the payload than Dataset 3, they were selected for this approach of training. The number of sequences per class was 162.

It was observed that there is indeed a lack of information in the sequences selected, particularly if only sequence features are included (Figure 8.4). The achieved accuracy rate after evaluation through 10-fold was 43% with only sequence features included, however, that increased when the combination of secondary structure and conservation features was thrown in, reporting around 62% accuracy (Table 8.1). This was actually better than Dataset 2, which had only the flanking region included (regardless of whether that overlapped an exon), that reported 51% accuracy. However, the rate is still well below the accuracy obtained from the flanking regions with more than 70%, using the combination of secondary structure information and conservation scores. Further performance metrics are available in the Appendix (Table A.8). Figure 8.5 shows the performance of the different models in cross-validation.

**Table 8.1:** 10-fold cross-validation results. Overview of 10-fold cross-validation accuracy for supervised machine learning on combinations of data and feature sets ($k$-mers only, $k$-mers plus secondary structure, or $k$-mers plus secondary structure and sequence conservation), both with and without the Fickett score as measure of coding potential. Sample sizes are indicative of the original dataset sizes, however, all classes were down-sampled to match the class with the smallest number of samples for a balanced classification task.

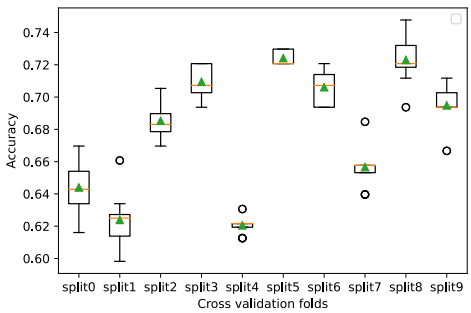
| Dataset | Fickett Score | $k$-mer | +structure | +conservation |
|---|---|---|---|---|
| Dataset 1 | | | | |
| sno: 345 | without | 85.50% | 85.51% | 84.06% |
| mir: 400 | with | 84.06% | 80.19% | 82.61% |
| lnc: 400 | | | | |
| Dataset 2 | | | | |
| sno: 690 | without | 44.69% | 47.58% | 51.69% |
| mir: 800 | with | 40.34% | 43.96% | 50.72% |
| lnc: 750 | | | | |
| Dataset 3 | | | | |
| sno:1287 | without | 66.67% | 67.03% | 70.61% |
| mir: 464 | with | 67.38% | 65.59% | 70.25% |
| lnc: 2101 | | | | |
| Dataset 4 | | | | |
| sno: 162 | without | 42.86% | 50.00% | 62.24% |
| mir: 168 | with | 41.84% | 35.71% | 50.00% |
| lnc: 750 | | | | |



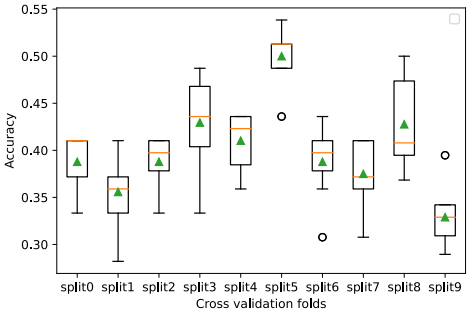

**(a)** Feature set 1




**(b)** Feature set 2




**(c)** Feature set 3




**(d)** Feature set 4

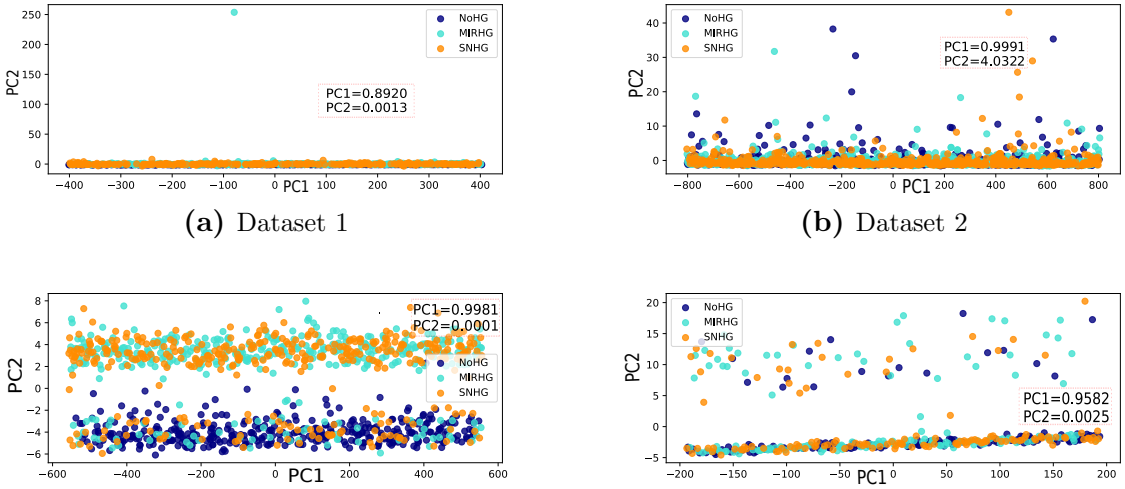
**Figure 8.5:** Cross-validation models. This figure shows four box plots showing the performances of the 10 models of the best estimators in the cross-validation strategies based upon the sequence features.

### 8.1.5 Unsupervised clustering

As an alternative approach to detecting commonalities within the three lncRNA classes, an unsupervised clustering approach was also developed motivated by the work of Kirk et al.[288], who presented a link between $k$-mer profiles and lncRNA function. An initial principal component analysis (PCA) of the four datasets based upon two principal components, however, did not reveal any credible clustering or separation between SNHG, MIRHG, and NoHG. An overview can be seen in 8.6.

Furthermore, employing $k$-means clustering it was noted that the assignment of the three lncRNA groups to the clusters is effectively random (Figure 8.7). Accuracy levels for all combinations and datasets are below 36%, moreover, the groups were always clustered around a single centroid. A detailed result of some of the other attributes of $k$-means clustering in available in the Appendix (Table A.2).

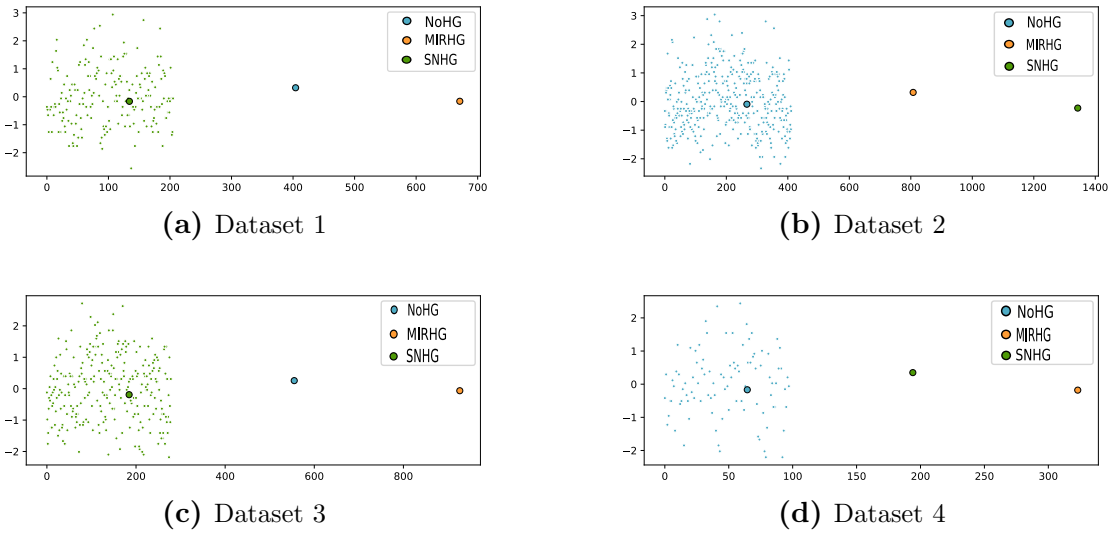
Following that, a convolutional neural network (CNN), a very effective approach in image classification, was designed in an attempt to discern any patterns within the data. The CNN could not show any credible distinctions between the lncRNA classes either. However, this was not unexpected, given the small size of the available data, since neural networks employed in deep learning methods always require massive amounts of data to learn from for a classification task.




**(a)** Dataset 1 **(b)** Dataset 2

**(c)** Dataset 3 **(d)** Dataset 4

**Figure 8.6:** Results of PCA. The overview of PCA carried out on the datasets giving the variance of the two principal components.




**(a)** Dataset 1 **(b)** Dataset 2

**(c)** Dataset 3 **(d)** Dataset 4

**Figure 8.7:** An overview of kmeans clustering carried out on the datasets. The classes always clustered around a single centroid, possibly referring to the lack of variance in the data.

# IV

# Discussion

# 9

# Discussion and conclusion

## 9.1 Lack of signals

The motivation behind the research work outlined in the previous chapter (Chapter 8) was to look for whether there is a clear distinction between the host genes of snoRNAs and miRNAs on one hand, and lncRNAs without such highly conserved ncRNAs as payloads on the other hand. The conclusion reached was largely negative. While machine learning methods readily distinguish the three classes based on generic features of their respective payloads (or the absence of a miRNA or snoRNA payload, respectively) mostly as a binary classification problem, the classification task seemingly becomes complicated, if the information about the payload and its vicinal regions is not made available to the classifier. While the sequence adjacent to the payload, in particular in the exonic part of the transcripts, appears to contain some payload-specific information, it was observed that the association between payload and features obtained from distant regions of the lncRNA is weak. Since a definite structure can be associated to segments of lncRNAs and parts of the sequences are also conserved, the features derived from secondary structures and conservation of these regions have in all cases a positive effect on classification accuracy and are readily ranked among the most important features. However, they do not entirely compensate for the weakness of the association with sequence-based features (Appendix Table A.3).

Applying $k$-means clustering to the datasets and features resulted in poor classification accuracy (Figure 8.7). Even though lncRNAs as such may harbour sequence motifs that give away conserved RNA binding protein target function among lncRNA groups, poor classification potential for MIRHGs and SNHGs was revealed in this study, even if the payload is considered. No distinctive clusters of MIRHGs and SNHGs were observed, rather all the sequences were gathered together. In this respect, these results are consistent with clustering of lncRNA classes observed in[288]. Many prominent SNHGs,

such as GAS5, have remained outside the identifiable clusters in that study.

To summarise, the computational experiments indicate that a credible distinction of SNHGs, MIRHGs, and NoHGs can be achieved with reasonable accuracy only based on the payload information, at least in case a supervised approach to training the classifier (Table 8.1). The flanking sequences arguably involved in the processing also convey some pertinent information, especially the adjacent exonic fragments, although it was observed that the attainable classification accuracy levels came back to be much smaller. If no payload related sequence information is included, the three classes become virtually indistinguishable, as can be noticed in the case where only distant exonic regions were included. However, unsupervised clustering methods appear to be unable to classify lncRNAs when constrained to taking only hosted RNAs into consideration, at least with the data available at the present. It appears that the signal that can be obtained from the smaller snoRNAs or miRNAs is drowned out by the differences among the much larger hosts. This at least suggests that there are no strong features that identify MIRHGs or SNHGs as coherent subgroups in lncRNAs.

## 9.2 Influence of the feature sets

The features designed for the classification task were motivated by their biological significance. The purely sequence based feature set of $k$-mers can be well interpreted as they encode meaningful information in short sequences[288]. The inclusion of predicted secondary structure and conservation scores served as a small benefit, especially for the sequence sets excluding the payload, but they cannot overcome the inability to train a classifier accurately unless the payload itself in included. Some of the features will be discussed in the following part.

### 9.2.1 Fickett score

It was invented as a measure of coding potential[280] and has been used diligently in machine learning problems as a feature to distinguish between protein-coding and non-coding RNAs[281;282]. Naturally, it was incorporated into this study as well to explore whether it could be benefitted from. Fickett score was always used as a switch with the feature sets; predictions were made with it both enabled and disabled. While a tiny increase in accuracy for flanking exons in Dataset 3 was observed when only sequence based features were used, none of the other predictions gained from its inclusion. Thus, it can be stated that the Fickett score does not constitute a meaningful feature for the classification problem at hand.

### 9.2.2 RNA secondary structure

With their double-stranded structures effectively used in various biological roles, snoRNAs fold meticulously and depend on them to be functional. Although miRNAs,

in their post-transcriptional regulatory capacity act, as a single-stranded entity when loaded into the RISC complex, prior to that phase in their biogenesis they are also heavily structured. Hence it was decided that secondary structures, or lack thereof, could indicate genomic regions which are better suited for the integration of either payload, *i. e.* snoRNAs or miRNAs. Using RNAplFold[299] of the ViennaRNA package[162;424], the probability of two nucleotides being unpaired was included as a feature vector into the prediction machinery, as explained in section 7.4.2 of Chapter 7. In contrast to the Fickett score, inclusion of these features always increased accuracy, although in most cases the improvement is small. Moreover, the combination of secondary structure features with Fickett score always resulted in a decreased prediction accuracy. Cumulatively, secondary structure features are informative for the classification task at hand, although their impact appears to be smaller than it was expected, given the abundant recorded observations of the structures of the genes and their efficacy in associated regulatory functions. A reason behind this lack of predictive power could be the apparent lack of conserved secondary structure of lncRNAs, which is otherwise found in plenty throughout the human transcriptome[425].

### 9.2.3 Sequence conservation

Sequence conservation is a key indicator of biological function. It therefore seemed sensible to include conservation features in our classification. However, it limited the application to genomic regions where reliable sequence alignments could be constructed - from which conservation scores could be deducted. The phastCons scores were obtained for every sequence in all the datasets and each individual score paired with each nucleotide of the candidate (sub-)sequences[303]. The scores provided a probability of a particular nucleotide being part of a conserved genomic region, as covered in section 7.4.2. Similar to secondary structure features, addition of conservation scores always increased accuracy values, except when the payload directly was considered in Dataset 1. The effect is actually the strongest when randomic exonic regions in Dataset 4 and could hint towards a general trend of conservation differences between the three classes of lncRNAs investigated.

### 9.2.4 miRNA target sites as a feature

Although miRNA target sites were not actually used as a feature in the classification task, the idea was explored, since certain SNHGs can act as miRNA sponges[328;329]. It could be hypothesised that these lncRNAs should be recognisable using the distribution of miRNA binding sites as features for classification, should that be the primary function of the exonic regions of the SNHGs. To test this hypothesis, predicted and experimentally validated miRNA binding sites from miRTarBase were retrieved[332], which contains not only target sites located in the 3' UTR, but also experimentally validated binding sites regardless of their genomic location, in contrast to other sources. However, intersection of this resource with the host genes considered for this study revealed no significant overlap of miRNA binding sites in any of the host lncRNAs. The datasets used for the classification task were populated with overlapping lncRNA genes

with precursors of snoRNA and miRNA genes. Since less than 0.1% of the reported miRNA target genes are lncRNAs, the complete list of seed regions for each miRNA available from TargetScan[426] was fetched and used and trained a classifier using only the $k$-mers appearing in these sequences. The search was restricted to regions covered in the analysed datasets, for which standardised, weighted $k$-mers were available. No conclusive enrichment for any of the analysed seeds could be detected when comparing SNHGs and the other classes. This approach is, of course, limited due to the rather small regions that was covered in the datasets.

## 9.3 Conclusion and remaining challenges

Annotation of genes has been challenging, as can be noticed when, for example, genomic loci in GENCODE collapse into a single one, merge with other existing loci, or disappear altogether across releases. Despite making leaps in progress, annotation strategies that are being employed presently are insufficient to be able to classify transcripts precisely, especially in the case of lncRNAs. There are several discrepancies between annotation databases, for example, between GENCODE and NONCODE, where the latter has manifold more transcripts listed than the former. Catalogues sharing the same annotation base (such as GENCODE and Ensembl) agree more. This suggests a fragmentation in the whole lncRNA annotation procedure. Iyer et al. reported[77] that earlier annotation strategies were mostly focussed upon detecting multi-exonic and intergenic transcripts, due to complex transcriptional reconstruction of the monoexonic or genic regions, which might have led to gaps in the annotated transcriptome. However, the approaches are changing progressively with the assistance of high-throughput sequencing technology that is available. Comparatively more accurate analysis of the data is also possible with access to various computational packages and tools.

At the end of studying splice junctions, it was seen that human transcriptome data harbour a large number of rare exons (and thus also introns) that have remained unannotated. Due to their low abundance, they appear only when data from large scale experiments are pooled. As shown in Fig. 5.2, they nevertheless can be reproduced very accurately. There is very little noise in these data, as shown by the near perfect saturation of the average number of splice junctions per gene. Transcriptional noise, whether biological or technical, would result in a linear increase of the number of detected junctions as function of the size of the data set. If such a slope exists, it is too small to be detectable from the lymphome data used, which comprise of more than $10^{10}$ reads. Therefore, it is to be concluded that the current annotation of the human transcriptome is confined to a very well defined set of splice variants. Concomitantly, it is a meaningful and a worthwhile task to attempt the construction of an exhaustive and comprehensive lncRNA classification and annotation atlas. The fact that the isoforms are well-defined does not does not automatically imply that all isoforms are carriers of biological functions. If the vast majority of isoforms are indeed non-functional junk, however, an explanation is needed for the precision of the processing and its restriction to very few splice sites.

The following study was about finding a way to distinguish snoRNA and miRNA host

genes without making the small RNAs accessible to the classifier, and, by extension, pick up signals hinting about the function of the host lncRNAs. The biological functions of the small RNAs are quite well-studied and that could pave a way to further understanding about lncRNA functions, should any relationship between the functions exist. Although a function as miRNA sponge has been reported for many SNHGs, features that might connect the sequence or structure of the SNHGs with specific $k$-mers (namely those complementary to the seed regions of miRNAs) or to predicted miRNA target sites could not be. Restricting the data to the small number of experimentally validated miRNA targets only severely limits the power of this feature, since only a very small fraction of known target genes are lncRNAs[427]. The use of predicted target sites, on the other hand, may suffer from high noise levels in the miRNA target predictions. A systematic investigation into miRNA targets on lncRNAs is still missing. Such an endeavor would help to shed light onto the regulatory interplay between SNHGs, MIRHGs and their payload.

From an evolutionary point of view, it may not be surprising that the host genes of miRNAs and snoRNAs do not exhibit recognisable class-specific features. Most likely, the molecular function of miRNAs and, in particular, snoRNAs is much older and predates functions of the non-coding host genes. These most likely arose secondarily, maybe long after the transcripts have come under negative selection as host genes. The lack of common, class-specific features for host genes together with their usually very poor sequence conservation suggests they may even have acquired different functions in different lineages. A better understanding of the host genes thus will require a much more detailed investigation into the patterns of conservation than what is available at present. While the evolution of snoRNAs and miRNAs has been received considerable attention[330;428–431], there are no systematic data on the conservation and evolutionary flexibility of their host lncRNAs.

Given that there is mounting evidence for biological function of not only lncRNAs but also specifically for snoRNA and miRNA host genes, this *lack of detectable association* can be confirmed to be of biological interest. It suggests that the function of the host genes is not closely tied to the function of the payload. This is in stark contrast to the protein-coding host genes of snoRNAs, many of which encode ribosomal proteins[372] and thus also contribute to the maturation of the ribosome. This can also be extended to lncRNAs, which originate from the overlapping regions or antisense strands or even promoter regions of other genes, and participate in the regulation of the host genes or the same locus.

SNHGs and MIRHGs are an excellent system to study the functional evolution of lncRNAs because the conserved payload makes it comparably easier to trace them over much larger evolutionary time scales than most other lncRNAs[432]. At the same time, both classes are large enough for statistical and learning based approaches. With rise of new sequencing technologies and advances in functional screening methods we can expect that more detailed data on host gene functions will be forthcoming. As it can be seen, molecular biology and genomics driven approaches juxtapose well and can be used to classify lncRNAs to better understand the intricacies of the transcriptome. More extensive functional data may also revise the picture of distribution of lncRNA functions, which shows only a rather loose association of biological function and molecular mechanism with sequence and structure features of transcripts. That overview

will certainly become effective for further research and applications.

Another aspect that can be noted is the classification of lncRNAs using machine learning techniques. There are numerous tools that perform this particular task and a subset of them have been explored in Chapter 6. Most of the tools have a common subset of features and separate lncRNAs from protein-coding transcripts well. The subset generally includes sequence-derived features, such as information on the ORF (length and coverage), transcript and mean exon lengths, and conservation scores of the nucleotides. A handful of tools include other features, such as secondary structure information and physiochemical properties. $k$-mer information is also utilised at times. The predominantly binary classification tasks of separating coding and non-coding genes seem to be served well by these features not only in humans, but also in other vertebrates, primarily in mammals. Although sequence conservation decreases with evolutionary distance[160], the other properties appear not to diminish and still play a role in establishing a lncRNA from an mRNA. The classification strategies take a mixed approach while training the classifiers. Some are trained on one species and validated with others, while the rest train the classifiers on individual species and attempt to perform the task of distinguish the two classes. The first situation is true in most cases; training on human transcripts and performing the classification on other mammals.

However, not a lot can be said about lncRNAs in plants. The transcription machinery is different in plants; the plant lncRNAs are transcribed by plant specific RNA pol-IV and pol-V, sometimes by pol-II, different from human lncRNAs[118]. Otherwise, they have been reported to have almost the same gene regulatory functions as human lncRNAs, such as acting as a decoy and recruitment of chromatin-modifying proteins, among others. Some transcripts transcribed by pol-IV undergo cleavage by Dicer-like 3 to form smaller siRNAs, which in turn are loaded onto AGO4 to interact with pol-V, leading to DNA methylation[118]. This is a regulatory function that is markedly different from human, or animal, lncRNAs[433]. From a computational point of view, the sequence features have been observed to be different from human transcripts by the authors of CNCI[296], when they performed the prediction on a plant dataset using a model trained on human lncRNAs. Its successor CNIT[297], however, performed better on plant datasets, the reason being the model was trained on *A. thaliana*. A tool designed specifically to classify plant-based coding and long non-coding transcripts, PLncPRO[434], also used sequence based features for the model training. The authors reported that it outperformed other existing tools on both plant-based and human datasets, although it should also be taken into account that the human dataset was predicted by a model trained on human transcripts. In this regard, it can be conjectured that as more knowledge about the roles and functions of plant lncRNAs are revealed, the underlying relationships with human (or animal) genes will be unveiled. Perhaps, only then it will be possible to better identify the dynamics between the two groups of lncRNAs and implement optimal features to be able to train a model in one species, e.g. human, and classify accurately in plants. That would also increase the general understanding of functional roles of lncRNAs.

# V

# Appendix

# A

# Additional Tables

**Table A.1:** Disease association of lncRNAs. LncRNAs have been implicated in many diseases, several cancer types being among the most studied. This table aims to collect a few of the disease-associated lncRNAs together along with vital regulatory mechanisms. N/A stands for unknown regulatory mechanisms

| Associated lncRNA | Disease | Regulatory function | Refs |
|---|---|---|---|
| HOTAIR | Esophageal, lung, cervical, pancreatic, breast, oral, hepatocellular and colorectal cancer | Chromatin modification | 435–442 |
| H19 | Bladder, gastric, esophageal, colorectal cancer and glioma, Beckwith-Wiedemann and Silver-Russell syndromes | PRC2 and LSD1 interaction, repression of suppressor genes | 178–181;183;443;444 |
| MALAT1 | Lung, colorectal, prostate, hepatocellular, uterine cancer and glioma | RNA splicing and protein interaction | 137;177;184;185;442;445 |
| HULC | Hepatocellular, pancreatic cancer and glioma | miRNA decoy | 177;446 |
| NEAT1 | Glioma, oral, hepatocellular and nasopharyngeal cancer | SWI/SNF interaction and repression | 177;189;447;448 |
| PVT1 | Thyroid, gastric and colorectal cancer, diabetes mellitus | EZH2 recruitment and TSHR regulation | 194;449–451 |
| ANRIL | Lung, hepatocellular and bladder cancer | Transcription regulation and gene silencing | 452–454 |
| PCAT-1 | Colorectal and prostate cancer | MYC regulation | 455 |
| CCAT1-L | Colorectal cancer | Promotes MYC transcription | 456 |
| CCAT2 | Lung, colon and cervical cancer | Destabilises chromosomal structure | 457–459 |
| UCA1 | Lung, oral, bladder, colon, hepatocellular, breast and esophageal cancer | miRNA sponge | 177;460 |
| AFAP1-AS1 | Lung, colorectal, hepatocellular and esophageal | Expression regulation | 461;462 |
| BC200 | Breast, cervical, esophageal, lung, ovary, parotid and tongue tumour | Protein interaction | 142;463;464 |
| GAS5 | Glioma and breast cancer | Glucocorticoid receptor decoy | 139 |
| PTENP1 | Gastric, breast and prostate cancer | miRNA decoy | 465 |
| SPRY4-IT1 | Melanoma and glioma | Protein interaction | 466 |
| KcnQ1ot1 | Breast, lung and colon cancer, Beckwith-Wiedemann syndrome | Epigenetic regulation | 444;467–469 |

| Associated lncRNA | Disease | Regulatory function | Refs |
|---|---|---|---|
| TUG1 | Glioma | Promotes cell apoptosis | 470 |
| MEG3 | Glioma, brain tumour | Represses p53 pathway in glioma | 471 |
| ZFAS1 | Breast, colorectal, oral and ovarian cancer | N/A | 472 |
| Wt1-as | Acute myeloid leukemia, breast and lung cancer, glioma | Interaction with sense WT1 with alternatively spliced isoforms | 473 |
| SChLAP1 | Prostate cancer | SWI/SNF interaction and repression | 188 |
| THCAT126 | Thyroid cancer | N/A | 77 |
| BRCAT49 | Breast cancer | N/A | 77 |
| MEAT6 | Melanoma | N/A | 77 |
| BACE1-AS | Alzheimer's disease | Stabilising mRNA | 190 |
| HYMAI | Transient neonatal diabetes mellitus | N/A | 474 |
| SNHG1 | Parkinson's Disease | miRNA regulation | 193 |
| lincRNA-Cox2 | Parkinson's Disease | Binds to NF-$\kappa$B p65 | 193 |
| MIAT | Myocardial infarction | N/A | 475 |
| PAXIP1-AP1 | Pulmonary Hypertension | N/A | 475 |
| CHRF | Heart failure | miRNA sponge | 194 |
| NONRATT021972 | Diabetes mellitus | N/A | 476 |

**Table A.2:** Results for $k$-means clustering. The *Training Accuracy* column shows the accuracy on the training set, with the count of accurately predicted labels in the *Correctly predicted* column. The frequencies of each class in both the training set and the test set are shown in the top row in the following two columns, respectively, with the actual predicted labels in the bottom row. The datasets were split 80-20.

| | Training Accuracy | Correctly predicted (Total) | Labels in trainingset (lnc, mir, sno)<br>Predicted labels [lnc, mir, sno] | Labels in testset (lnc, mir, sno)<br>Predicted labels [lnc, mir, sno] |
|---|---|---|---|---|
| Dataset 1 | 34.42% | 285 (828) | (282, 277, 269)<br>[278,272,278] | (63, 68, 76)<br>[0, 0, 207] |
| Dataset 2 | 33.33% | 579 (1656) | (548, 563, 545)<br>[549,557,550] | (142,127,145)<br>[414, 0, 0] |
| Dataset 3 | 35.76% | 398 (1113) | (371, 373, 369)<br>[371,372,370] | (93, 91, 95)<br>[0, 0, 279] |
| Dataset 4 | 34.79% | 135 (388) | (125, 136, 127)<br>[130, 129, 129] | (37, 26, 35)<br>[98, 0, 0] |

**Table A.3:** Feature importance. Five most important features for classification according to the random forest classifier. $x-z$-mers should be interpreted as all $k$-mers from length x through z.

| Dataset | Featureset | Fickett Score | Rank 1 | Rank 2 | Rank 3 | Rank 4 | Rank 5 |
|---|---|---|---|---|---|---|---|
| Dataset 1 | k-mer | without | (3-4mers) | (CG, 3-5mers) | (3-4mers) | (3-6mers) | (3-5mers) |
| | | with | (2-4mers) | (2mers, 4-5mers) | (3-4mers) | (3-4mers) | (4-5mers) |
| | +structure | without | (4-5mers) | (4-5mers) | (4-5mers) | (4-5mers) | (4mers) |
| | | with | (3-4mers) | (4mers) | (4-5mers) | (4-6mers) | (4-6mers) |
| | +conservation | without | (CCG, 4-6mers) | (3-4mers, AGCGG) | (CG, 3-5mers) | (CG, 3-7mers) | (3-7mers) |
| | | with | (3-4mers, AGCGG) | (CG, 3-7mers) | (CGA, 4-5mers, 7mers) | (3-4mers, CAGCG) | (CG, 3-7mers) |
| Dataset 2 | k-mer | without | GC content | Sequence | (2mers) | (2mers) | (2mers) |
| | | with | Sequence | GC content | Fickett score | (2mers) | (2mers) |
| | +structure | without | GC content | Structure | Structure | Sequence | Structure |
| | | with | GC content | Structure | Fickett score | Structure | Structure |
| | +conservation | without | Conservation | GC content | Sequence | Conservation | Structure |
| | | with | Conservation | GC content | Conservation | Structure | Sequence |
| Dataset 3 | k-mer | without | GC content | (7mers) | Sequence | (3-4mers) | (3-4mers) |
| | | with | GC content | (7mers) | Fickett score | Sequence | (7mers) |
| | +structure | without | GC content | (7mers) | Structure | Structure | Structure |
| | | with | GC content | (7mers) | (7mers) | Structure | Structure |
| | +conservation | without | GC content | (7mers) | (3-4mers) | Structure | Structure |
| | | with | (7mers) | GC content | Structure | (7mers) | Structure |
| Dataset 4 | k-mer | without | GC content | Sequence | (2mers) | (2mers) | (2-3mers, AAAAA, AAAAAA) |
| | | with | GC content | Fickett score | Sequence | (2mers) | (2-3mers, AAAC, 5-6mers) |
| | +structure | without | GC content | Structure | Structure | Structure | Structure |
| | | with | Structure | GC content | Structure | Structure | Sequence |
| | +conservation | without | GC content | Structure | Sequence | Structure | Structure |
| | | with | GC content | Structure | Structure | Structure | Structure |

**Table A.4:** Hyperparameter settings. Accuracy after hyperparameter tuning by grid search using 10-fold cross-validation metric for supervised machine learning. Indicated are accuracy per feature set and parameters for each feature set as [criterion, number of trees]. Bootstrap and out-of-bag score were always enabled. F stands for Fickett score.

| Dataset | F | kmer | | +structure | | +conservation | |
|---|---|---|---|---|---|---|---|
| Dataset 1 | | | | | | | |
| SNHGs: 345 | without | 85.50% | [‘gini’, 300] | 85.51% | [‘entropy’, 100] | 84.06% | [‘entropy’, 1000] |
| MIRHGs: 400 | with | 84.06% | [‘gini’, 500] | 80.19% | [‘gini’, 500] | 82.61% | [‘entropy’, 100] |
| NoHGs: 400 | | | | | | | |
| Dataset 2 | | | | | | | |
| SNHGs: 690 | without | 44.69% | [‘entropy’, 1000] | 47.58% | [‘entropy’, 1000] | 51.69% | [‘entropy’, 100] |
| MIRHGs: 800 | with | 40.34% | [‘entropy’, 300] | 43.96% | [‘entropy’, 500] | 50.72% | [‘entropy’, 1000] |
| NoHGs: 750 | | | | | | | |
| Dataset 3 | | | | | | | |
| SNHGs: 1287 | without | 66.67% | [‘gini’, 1000] | 67.03% | [‘entropy’, 500] | 70.61% | [‘gini’, 100] |
| MIRHGs: 464 | with | 67.38% | [‘entropy’, 100] | 65.59% | [‘entropy’, 500] | 70.25% | [‘gini’, 300] |
| NoHGs: 2101 | | | | | | | |
| Dataset 4 | | | | | | | |
| SNHGs: 162 | without | 42.86% | [‘entropy’, 1000] | 50.00% | [‘gini’, 100] | 62.24% | [‘gini’, 500] |
| MIRHGs: 168 | with | 41.84% | [‘entropy’, 1000] | 35.71% | [‘entropy’, 100] | 50.00% | [‘entropy’, 100] |
| NoHGs: 750 | | | | | | | |

**Table A.5:** Performance metrics for Dataset 1. The metrics for Dataset 1 without Fickett score. The support column indicates the number of elements for each class in the test set.

| Feature Set | Class | precision | recall | f1-score | support |
|---|---|---|---|---|---|
| 1 | NoHG | 0.95 | 1.00 | 0.98 | 63 |
| | MIRHG | 0.77 | 0.81 | 0.79 | 68 |
| | SNHG | 0.84 | 0.78 | 0.81 | 76 |
| 2 | NoHG | 0.95 | 1.00 | 0.98 | 63 |
| | MIRHG | 0.72 | 0.87 | 0.79 | 68 |
| | SNHG | 0.88 | 0.68 | 0.77 | 76 |
| 3 | NoHG | 0.95 | 1.00 | 0.98 | 63 |
| | MIRHG | 0.76 | 0.81 | 0.79 | 68 |
| | SNHG | 0.84 | 0.76 | 0.80 | 76 |

**Table A.6:** The metrics for Dataset 2 without Fickett score. The support column indicates the number of elements for each class in the test set.

| Feature Set | Class | precision | recall | f1-score | support |
|---|---|---|---|---|---|
| 1 | NoHG | 0.39 | 0.30 | 0.34 | 142 |
| | MIRHG | 0.38 | 0.55 | 0.45 | 127 |
| | SNHG | 0.49 | 0.41 | 0.44 | 145 |
| 2 | NoHG | 0.43 | 0.37 | 0.39 | 142 |
| | MIRHG | 0.38 | 0.45 | 0.41 | 127 |
| | SNHG | 0.49 | 0.48 | 0.48 | 145 |
| 3 | NoHG | 0.51 | 0.59 | 0.55 | 142 |
| | MIRHG | 0.48 | 0.47 | 0.48 | 127 |
| | SNHG | 0.52 | 0.44 | 0.48 | 145 |

**Table A.7:** The metrics for Dataset 3 without Fickett score. The support column indicates the number of elements for each class in the test set.

| Feature Set | Class | precision | recall | f1-score | support |
|---|---|---|---|---|---|
| 1 | NoHG | 0.65 | 0.99 | 0.78 | 93 |
| | MIRHG | 0.69 | 0.49 | 0.58 | 91 |
| | SNHG | 0.65 | 0.49 | 0.56 | 95 |
| 2 | NoHG | 0.64 | 1.00 | 0.78 | 93 |
| | MIRHG | 0.69 | 0.57 | 0.63 | 91 |
| | SNHG | 0.71 | 0.43 | 0.54 | 95 |
| 3 | NoHG | 0.65 | 1.00 | 0.78 | 93 |
| | MIRHG | 0.76 | 0.60 | 0.67 | 91 |
| | SNHG | 0.78 | 0.52 | 0.62 | 95 |

**Table A.8:** The metrics for Dataset 4 without Fickett score. The support column indicates the number of elements for each class in the test set.

| Feature Set | Class | precision | recall | f1-score | support |
|---|---|---|---|---|---|
| 1 | NoHG | 0.42 | 0.41 | 0.41 | 37 |
| | MIRHG | 0.38 | 0.42 | 0.40 | 26 |
| | SNHG | 0.39 | 0.37 | 0.38 | 35 |
| 2 | NoHG | 0.54 | 0.38 | 0.44 | 37 |
| | MIRHG | 0.38 | 0.46 | 0.41 | 26 |
| | SNHG | 0.57 | 0.66 | 0.61 | 35 |
| 3 | NoHG | 0.67 | 0.65 | 0.66 | 37 |
| | MIRHG | 0.48 | 0.38 | 0.43 | 26 |
| | SNHG | 0.56 | 0.66 | 0.61 | 35 |

# Bibliography


[1] Run-Wen Yao, Yang Wang, and Ling-Ling Chen. Cellular functions of long noncoding RNAs. *Nature cell biology*, 21(5):542–551, 2019. doi:10.1038/s41556-019-0311-8.

[2] Mitchell Guttman and John L Rinn. Modular regulatory principles of large non-coding RNAs. *Nature*, 482(7385):339–346, 2012. doi:10.1038/nature10887.

[3] E. K. Robinson, S. Covarrubias, and S. Carpenter. The how and why of lncRNA function: An innate immune perspective. *Biochim Biophys Acta Gene Regul Mech*, 1863(4):194419, 04 2020. doi:10.1016/j.bbagrm.2019.194419.

[4] Barbara Uszczynska-Ratajczak, Julien Lagarde, Adam Frankish, Roderic Guigó, and Rory Johnson. Towards a complete map of the human long non-coding RNA transcriptome. *Nature Reviews Genetics*, 19(9):535–548, 2018. doi:10.1038/s41576-018-0017-y.

[5] Adam Frankish, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, Cristina Sisu, James Wright, Joel Armstrong, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research*, 47(D1):D766–D773, 2019. doi:10.1093/nar/gky955.

[6] Julia Richter, Matthias Schlesner, Steve Hoffmann, Markus Kreuz, Ellen Leich, Birgit Burkhardt, Maciej Rosolowski, Ole Ammerpohl, Rabea Wagener, Stephan H Bernhart, et al. Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nature genetics*, 44(12):1316, 2012. doi:10.1038/ng.2469.

[7] Michelle S Scott and Motoharu Ono. From snoRNA to miRNA: Dual function regulatory non-coding RNAs. *Biochimie*, 93(11):1987–1992, 2011. doi:10.1016/j.biochi.2011.05.026.

[8] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 2001. doi:10.1038/35057062.

[9] ENCODE Project Consortium et al. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004.

[10] ENCODE Project Consortium et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799, 2007. doi:10.1126/science.1105136)(PMID:15499007).

[11] ShuangSang Fang, LiLi Zhang, JinCheng Guo, YiWei Niu, Yang Wu, Hui Li, LianHe Zhao, XiYuan Li, XueYi Teng, XianHui Sun, et al. NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic acids research*, 46(D1):D308–D314, 2018. doi:10.1093/nar/gkx1107.

[12] Irina Maljkovic Berry, Melanie C Melendrez, Kimberly A Bishop-Lilly, Wiriya Rutvisuttinunt, Simon Pollett, Eldin Talundzic, Lindsay Morton, and Richard G Jarman. Next Generation Sequencing and Bioinformatics Methodologies for Infectious Disease Research and Public Health: Approaches, Applications, and Considerations for Development of Laboratory Capacity. *The Journal of Infectious Diseases*, 221(Supplement_3):S292–S307, 2020. doi:10.1093/infdis/jiz286.

[13] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6):333, 2016. doi:10.1038/nrg.2016.49.

[14] Elaine R Mardis. Next-generation sequencing platforms. *Annual review of analytical chemistry*, 6:287–303, 2013. doi:10.1146/annurev-anchem-062012-092628.

[15] Frederick Sanger, Steven Nicklen, and Alan R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12):5463–5467, 1977. doi:10.1073/pnas.74.12.5463.

[16] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008. doi:10.1038/nature07517.

[17] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, 475(7356):348–352, 2011. doi:10.1038/nature10242.

[18] Mark JP Chaisson, John Huddleston, Megan Y Dennis, Peter H Sudmant, Maika Malig, Fereydoun Hormozdiari, Francesca Antonacci, Urvashi Surti, Richard Sandstrom, Matthew Boitano, et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, 517(7536):608–611, 2015. doi:10.1038/nature13907.

[19] Kin Fai Au, Jason G Underwood, Lawrence Lee, and Wing Hung Wong. Improving PacBio long read accuracy by short read alignment. *PloS one*, 7(10):e46679, 2012. doi:10.1371/journal.pone.0046679.

[20] Rory Stark, Marta Grzelak, and James Hadfield. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019. doi:10.1038/s41576-019-0150-2.

[21] Miten Jain, Ian T Fiddes, Karen H Miga, Hugh E Olsen, Benedict Paten, and Mark Akeson. Improved data analysis for the MinION nanopore sequencer. *Nature methods*, 12(4):351–356, 2015. doi:10.1038/nmeth.3290.

[22] Sara Goodwin, James Gurtowski, Scott Ethe-Sayers, Panchajanya Deshpande, Michael C Schatz, and W Richard McCombie. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*, 25(11):1750–1756, 2015. doi:10.1101/gr.191395.115.

[23] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008. doi:10.1038/nature07509.

[24] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012. doi:10.1038/nature11233.

[25] Isabella Gazzoli, Irina Pulyakhina, Nisha E Verwey, Yavuz Ariyurek, Jeroen FJ Laros, Peter AC 't Hoen, and Annemieke Aartsma-Rus. Non-sequential and multi-step splicing of the dystrophin transcript. *RNA biology*, 13(3):290–305, 2016. doi:10.1080/15476286.2015.1125074.

[26] Pär G Engström, Tamara Steijger, Botond Sipos, Gregory R Grant, André Kahles, Tyler Alioto, Jonas Behr, Paul Bertone, Regina Bohnert, Davide Campagna, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature methods*, 10(12):1185–1191, 2013. doi:10.1038/nmeth.2722.

[27] Rachael E Workman, Alison D Tang, Paul S Tang, Miten Jain, John R Tyson, Roham Razaghi, Philip C Zuzarte, Timothy Gilpatrick, Alexander Payne, Joshua Quick, et al. Nanopore native RNA sequencing of a human poly (A) transcriptome. *Nature methods*, 16(12):1297–1305, 2019. doi:10.1038/s41592-019-0617-2.

[28] Daniel R Garalde, Elizabeth A Snell, Daniel Jachimowicz, Botond Sipos, Joseph H Lloyd, Mark Bruce, Nadia Pantic, Tigist Admassu, Phillip James, Anthony Warland, et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nature methods*, 15(3):201, 2018. doi:10.1038/nmeth.4577.

[29] Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, 6, 2017. doi:10.12688/f1000research.10571.2.

[30] Michael A Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13(1):1–13, 2012. doi:10.1186/1471-2164-13-341.

[31] Melissa J Fullwood, Chia-Lin Wei, Edison T Liu, and Yijun Ruan. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome research*, 19(4):521–532, 2009. doi:10.1101/gr.074906.107.

[32] Mitsuyoshi Murata, Hiromi Nishiyori-Sueki, Miki Kojima-Ishiyama, Piero Carninci, Yoshihide Hayashizaki, and Masayoshi Itoh. Detecting expressed genes using CAGE. In *Transcription Factor Regulatory Networks*, pages 67–85. Springer, 2014. doi:10.1007/978-1-4939-0805-9_7.

[33] Y. Zhang, X. Liu, and J. et al. MacLeod. Discerning novel splice junctions derived from RNA-seq alignment: a deep learning approach. *BMC Genomics*, 19(971), 2018. doi:10.1186/s12864-018-5350-1.

[34] Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*, 8(8):1494–1512, 2013. doi:10.1038/nprot.2013.084.

[35] Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, Martin Akerman, Tyler Alioto, Giovanna Ambrosini, Stylianos E Antonarakis, Jonas Behr, Paul Bertone, et al. Assessment of transcript reconstruction methods for RNA-seq. *Nature methods*, 10(12):1177–1184, 2013. doi:10.1038/nmeth.2714.

[36] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J Van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515, 2010. doi:10.1038/nbt.1621.

[37] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011. doi:10.1038/nbt.1883.

[38] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, 2013. doi:10.1186/gb-2013-14-4-r36.

[39] Steve Hoffmann, Christian Otto, Gero Doose, Andrea Tanzer, David Langenberger, Sabina Christ, Manfred Kunz, Lesca M Holdt, Daniel Teupser, Jörg Hackermüller, et al. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome biology*, 15(2):R34, 2014. doi:10.1186/gb-2014-15-2-r34.

[40] Qiong-Yi Zhao, Yi Wang, Yi-Meng Kong, Da Luo, Xuan Li, and Pei Hao. Optimizing de novo transcriptome assembly from short-read RNA-seq data: a comparative study. In *BMC bioinformatics*, volume 12, page S2. Springer, 2011. doi:10.1186/1471-2105-12-S14-S2.

[41] Thomas Gatter and Peter F Stadler. Ryuto: A Framework for Network-Flow based Transcriptome Reconstruction of RNA-seq Data. *BMC bioinformatics*, 20(9), 2019. doi:10.1186/s12859-019-2786-5.

[42] Peter J Shepard, Eun-A Choi, Jente Lu, Lisa A Flanagan, Klemens J Hertel, and Yongsheng Shi. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17 (4):761–772, 2011. doi:10.1261/rna.2581711.

[43] Dafne Campigli Di Giammartino, Kensei Nishida, and James L Manley. Mechanisms and consequences of alternative polyadenylation. *Molecular cell*, 43(6):853–866, 2011. doi:10.1016/j.molcel.2011.08.017.

[44] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature methods*, 5(7):621–628, 2008. doi:10.1038/nmeth.1226.

[45] Nagarjun Vijay, Jelmer W Poelstra, Axel Künstner, and Jochen BW Wolf. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular ecology*, 22(3):620–634, 2013. doi:10.1111/mec.12014.

[46] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*, 7(3): 562–578, 2012. doi:10.1038/nprot.2012.016.

[47] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature biotechnology*, 32(9):896–902, 2014. doi:10.1038/nbt.2931.

[48] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014. doi:10.1186/s13059-014-0550-8.

[49] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. Science forum: the human cell atlas. *eLife*, 6:e27041, 2017. doi:10.7554/eLife.27041.

[50] Thomas R Insel, Story C Landis, and Francis S Collins. The NIH brain initiative. *Science*, 340 (6133):687–688, 2013. doi:10.1126/science.1239276.

[51] Leighton J Core, Joshua J Waterfall, and John T Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848, 2008. doi:10.1126/science.1162228.

[52] Nicholas T Ingolia, Sina Ghaemmaghami, John RS Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223, 2009. doi:10.1126/science.1168978.

[53] Jesse M Engreitz, Klara Sirokman, Patrick McDonel, Alexander A Shishkin, Christine Surka, Pamela Russell, Sharon R Grossman, Amy Y Chow, Mitchell Guttman, and Eric S Lander. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. *Cell*, 159(1):188–199, 2014. doi:10.1016/j.cell.2014.08.018.

[54] David S Johnson, Ali Mortazavi, Richard M Myers, and Barbara Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, 2007. doi:10.1126/science.1141319.

[55] Aitor Garzia, Cindy Meyer, Pavel Morozov, Marcin Sajek, and Thomas Tuschl. Optimization of PAR-CLIP for transcriptome-wide identification of binding sites of RNA-binding proteins. *Methods*, 118:24–40, 2017. doi:10.1016/j.ymeth.2016.10.007.

[56] Thomas Derrien, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, David Martin, Angelika Merkel, David G Knowles, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9):1775–1789, 2012. doi:10.1101/gr.132159.111.

[57] Mitchell Guttman, Ido Amit, Manuel Garber, Courtney French, Michael F Lin, David Feldser, Maite Huarte, Or Zuk, Bryce W Carey, John P Cassady, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235):223–227, 2009. doi:10.1038/nature07672.

[58] H. Hezroni, D. Koppstein, M. G. Schwartz, A. Avrutin, D. P. Bartel, and I. Ulitsky. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep*, 11(7):1110–1122, May 2015. doi:10.1016/j.celrep.2015.04.023.

[59] A. Kanitz, F. Gypas, A. J. Gruber, A. R. Gruber, G. Martin, and M. Zavolan. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biol*, 16:150, Jul 2015. doi:10.1186/s13059-015-0702-5.

[60] S. J. Liu, M. A. Horlbeck, S. W. Cho, H. S. Birk, M. Malatesta, D. He, F. J. Attenello, J. E. Villalta, M. Y. Cho, Y. Chen, M. A. Mandegar, M. P. Olvera, L. A. Gilbert, B. R. Conklin, H. Y. Chang, J. S. Weissman, and D. A. Lim. CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, 355(6320), 01 2017. doi:10.1126/science.aah7111.

[61] H. Cao, C. Wahlestedt, and P. Kapranov. Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls. *Trends Genet*, 34(9):704–721, 09 2018. doi:10.1016/j.tig.2018.06.002.

[62] I. W. Deveson, M. E. Brunck, J. Blackburn, E. Tseng, T. Hon, T. A. Clark, M. B. Clark, J. Crawford, M. E. Dinger, L. K. Nielsen, J. S. Mattick, and T. R. Mercer. Universal Alternative Splicing of Noncoding Exons. *Cell Syst*, 6(2):245–255, Feb 2018. doi:10.1101/136275.

[63] S. A. Hardwick, W. Y. Chen, T. Wong, I. W. Deveson, J. Blackburn, S. B. Andersen, L. K. Nielsen, J. S. Mattick, and T. R. Mercer. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat Methods*, 13(9):792–798, 09 2016. doi:10.1038/nmeth.3958.

[64] D. Sharon, H. Tilgner, F. Grubert, and M. Snyder. A single-molecule long-read survey of the human transcriptome. *Nat Biotechnol*, 31(11):1009–1014, Nov 2013. doi:10.1038/nbt.2705.

[65] T. R. Mercer, D. J. Gerhardt, M. E. Dinger, J. Crawford, C. Trapnell, J. A. Jeddeloh, J. S. Mattick, and J. L. Rinn. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol*, 30(1):99–104, Nov 2011. doi:10.1038/nbt.2024.

[66] Rituparno Sen, Gero Doose, and Peter F Stadler. Rare splice variants in long non-coding RNAs. *Non-coding RNA*, 3(3):23, 2017. doi:https://doi.org/10.3390/ncrna3030023.

[67] Ahmad M. Khalil, Mitchell Guttman, Maite Huarte, Manuel Garber, Arjun Raj, Dianali Rivea Morales, Kelly Thomas, Aviva Presser, Bradley E. Bernstein, Alexander van Oudenaarden, Aviv Regev, Eric S. Lander, and John L. Rinn. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences*, 106(28):11667–11672, 2009. ISSN 0027-8424. doi:10.1073/pnas.0904715106.

[68] Julien Jarroux, Antonin Morillon, and Marina Pinskaya. History, discovery, and classification of lncRNAs. In *Long Non Coding RNA Biology*, pages 1–46. Springer, 2017. doi:10.1007/978-981-10-5203-3_1.

[69] Mohammad Ali Faghihi, Farzaneh Modarresi, Ahmad M Khalil, Douglas E Wood, Barbara G Sahagan, Todd E Morgan, Caleb E Finch, Georges St Laurent III, Paul J Kenny, and Claes Wahlestedt. Expression of a noncoding RNA is elevated in alzheimer's disease and drives rapid feed-forward regulation of $\beta$-secretase. *Nature medicine*, 14(7):723–730, 2008. doi:10.1038/nm1784.

[70] Claudia Carrieri, Laura Cimatti, Marta Biagioli, Anne Beugnet, Silvia Zucchelli, Stefania Fedele, Elisa Pesce, Isidre Ferrer, Licio Collavin, Claudio Santoro, et al. Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature*, 491(7424):454–457, 2012. doi:10.1038/nature11508.

[71] Qing-Fei Yin, Li Yang, Yang Zhang, Jian-Feng Xiang, Yue-Wei Wu, Gordon G Carmichael, and Ling-Ling Chen. Long noncoding RNAs with snoRNA ends. *Molecular cell*, 48(2):219–230, 2012. doi:10.1016/j.molcel.2012.07.033.

[72] Alireza Shahryari, Marie Saghaeian Jazi, Nader M Samaei, and Seyed J Mowla. Long non-coding RNA SOX2OT: expression signature, splicing patterns, and emerging roles in pluripotency and tumorigenesis. *Frontiers in genetics*, 6:196, 2015. doi:10.3389/fgene.2015.00196.

[73] Thomas B Hansen, Trine I Jensen, Bettina H Clausen, Jesper B Bramsen, Bente Finsen, Christian K Damgaard, and Jørgen Kjems. Natural RNA circles function as efficient microRNA sponges. *Nature*, 495(7441):384–388, 2013. doi:10.1038/nature11993.

[74] Suman Ghosal, Shaoli Das, Rituparno Sen, Piyali Basak, and Jayprokas Chakrabarti. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Frontiers in genetics*, 4:283, 2013. doi:10.3389/fgene.2013.00283.

[75] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox, O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1): D733–745, Jan 2016. doi:10.1093/nar/gkv1189.

[76] Chung-Chau Hon, Jordan A Ramilowski, Jayson Harshbarger, Nicolas Bertin, Owen JL Rackham, Julian Gough, Elena Denisenko, Sebastian Schmeier, Thomas M Poulsen, Jessica Severin, et al. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, 543(7644):199–204, 2017. doi:10.1038/nature21374.

[77] Matthew K Iyer, Yashar S Niknafs, Rohit Malik, Udit Singhal, Anirban Sahu, Yasuyuki Hosono, Terrence R Barrette, John R Prensner, Joseph R Evans, Shuang Zhao, et al. The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics*, 47(3):199–208, 2015. doi:10.1038/ng.3192.

[78] B. A. Sweeney, A. I. Petrov, C. E. Ribas, R. D. Finn, A. Bateman, M. Szymanski, W. M. Karlowski, S. E. Seemann, J. Gorodkin, J. J. Cannone, R. R. Gutell, S. Kay, S. Marygold, G. Dos Santos, A. Frankish, J. M. Mudge, R. Barshir, S. Fishilevich, P. P. Chan, T. M. Lowe, R. Seal, E. Bruford, S. Panni, P. Porras, D. Karagkouni, A. G. Hatzigeorgiou, L. Ma, Z. Zhang, P. J. Volders, P. Mestdagh, S. Griffiths-Jones, B. Fromm, K. J. Peterson, I. Kalvari, E. P. Nawrocki, A. S. Petrov, S. Weng, P. Bouchard-Bourelle, M. Scott, L. M. Lui, D. Hoksza, R. C. Lovering, B. Kramarz, P. Mani, S. Ramachandran, and Z. Weinberg. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic acids research*, Oct 2020. doi:10.1093/nar/gkaa921.

[79] L. Ma, A. Li, D. Zou, X. Xu, L. Xia, J. Yu, V. B. Bajic, and Z. Zhang. LncRNAWiki: harnessing community knowledge in collaborative curation of human long non-coding RNAs. *Nucleic acids research*, 43(Database issue):D187–192, Jan 2015. doi:10.1093/nar/gku1167.

[80] P. J. Volders, J. Anckaert, K. Verheggen, J. Nuytens, L. Martens, P. Mestdagh, and J. Vandesompele. LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic acids research*, 47(D1):D135–D139, 01 2019. doi:10.1093/nar/gky1031.

[81] Ralf Dahm. Friedrich Miescher and the discovery of DNA. *Developmental biology*, 278(2):274–288, 2005. doi:https://doi.org/10.1016/j.ydbio.2004.11.028.

[82] Oswald T Avery, Colin M MacLeod, and Maclyn McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *The Journal of experimental medicine*, 79(2):137–158, 1944.

[83] Matthew Cobb. Who discovered messenger RNA? *Current Biology*, 25(13):R526–R532, 2015. doi:10.1016/j.cub.2015.05.032.

[84] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[85] Francis HC Crick. The origin of the genetic code. *Journal of molecular biology*, 38(3):367–379, 1968.

[86] JB Lewis, JF Atkins, CW Anderson, PR Baum, and RF Gesteland. Mapping of late adenovirus genes by cell-free translation of RNA selected by hybridization to specific DNA fragments. *Proceedings of the National Academy of Sciences*, 72(4):1344–1348, 1975. doi:10.1073/pnas.72.4.1344.

[87] Arnold J Berk. Discovery of RNA splicing and genes in pieces. *Proceedings of the National Academy of Sciences*, 113(4):801–805, 2016. doi:10.1073/pnas.1525084113.

[88] Gary Zieve and Sheldon Penman. Small RNA species of the Hela cell: metabolism and subcellular localization. *Cell*, 8(1):19–31, 1976. doi:10.1016/0092-8674(76)90181-1.

[89] Thomas R Cech. The ribosome is a ribozyme. *Science*, 289(5481):878–879, 2000. doi:10.1126/science.289.5481.878.

[90] Stephane E Castel and Robert A Martienssen. Rna interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nature Reviews Genetics*, 14(2):100–112, 2013. doi:10.1038/nrg3355.

[91] Harold S Bernhardt. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others). *Biology direct*, 7(1):23, 2012. doi:10.1186/1745-6150-7-23.

[92] Masayori Inouye and Nicholas Delihast. Small RNAs in the prokaryotes: a growing list of diverse roles. *Cell*, 53(1):5–7, 1988. doi:10.1016/0092-8674(88)90480-1.

[93] Stefan L Ameres and Phillip D Zamore. Diversifying microRNA sequence and function. *Nature reviews Molecular cell biology*, 14(8):475–488, 2013. doi:10.1038/nrm3611.

[94] DP Barlow, R Stöger, BG Herrmann, K Saito, and N Schweifer. The mouse insulin-like growth factor type-2 receptor is imprinted and closely linked to the Tme locus. *Nature*, 349(6304):84–87, 1991. doi:10.1038/349084a0.

[95] Camilynn I Brannan, ELIZABETH CLAIRE Dees, Robert S Ingram, and SHIRLEY M Tilghman. The product of the H19 gene may function as an RNA. *Molecular and cellular biology*, 10(1): 28–36, 1990. doi:10.1128/mcb.10.1.28.

[96] Carolyn J Brown, Andrea Ballabio, James L Rupert, Ronald G Lafreniere, Markus Grompe, Rossana Tonlorenzi, and Huntington F Willard. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, 349(6304): 38–44, 1991. doi:10.1038/349038a0.

[97] Corinne Chureau, Marine Prissette, Agnès Bourdet, Valérie Barbe, Laurence Cattolico, Louis Jones, André Eggen, Philip Avner, and Laurent Duret. Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome research*, 12(6):894–908, 2002. doi:10.1101/gr.152902.

[98] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931, 2004. doi:10.1038/nature03001.

[99] Philipp Kapranov, Simon E Cawley, Jorg Drenkow, Stefan Bekiranov, Robert L Strausberg, Stephen PA Fodor, and Thomas R Gingeras. Large-scale transcriptional activity in chromosomes 21 and 22. *Science*, 296(5569):916–919, 2002. doi:10.1126/science.1068597.

[100] Pea Carninci, T Kasukawa, S Katayama, J Gough, MC Frith, Norihiro Maeda, Rieko Oyama, T Ravasi, B Lenhard, C Wells, et al. The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563, 2005. doi:10.1126/science.1112014.

[101] Ewan Birney. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489 (7414):57–74, 2012. doi:10.1038/nature11247.

[102] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James GR Gilbert, Roy Storey, David Swarbreck, et al. GENCODE: producing a reference annotation for ENCODE. *Genome biology*, 7(1):1–9, 2006. doi:10.1186/gb-2006-7-s1-s4.

[103] John S Mattick. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25(10):930–939, 2003. doi:10.1002/bies.10332.

[104] Moran N Cabili, Margaret C Dunagin, Patrick D McClanahan, Andrew Biaesch, Olivia Padovan-Merhar, Aviv Regev, John L Rinn, and Arjun Raj. Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome biology*, 16(1):20, 2015. doi:10.1186/s13059-015-0586-4.

[105] Radha Raman Pandey, Tanmoy Mondal, Faizaan Mohammad, Stefan Enroth, Lisa Redrup, Jan Komorowski, Takashi Nagano, Debora Mancini-DiNardo, and Chandrasekhar Kanduri. Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Molecular cell*, 32(2):232–246, 2008. doi:10.1016/j.molcel.2008.08.022.

[106] John L. Rinn, Michael Kertesz, Jordon K. Wang, Sharon L. Squazzo, Xiao Xu, Samantha A. Brugmann, L. Henry Goodnough, Jill A. Helms, Peggy J. Farnham, Eran Segal, and Howard Y. Chang. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311 – 1323, 2007. ISSN 0092-8674. doi:https://doi.org/10.1016/j.cell.2007.05.022.

[107] Tim R. Mercer, Marcel E. Dinger, and John S. Mattick. Long non-coding RNAs: insights into functions. *Nature Reviews Genetics*, 10(3):155–159, 2009. doi:10.1038/nrg2521.

[108] Maite Huarte, Mitchell Guttman, David Feldser, Manuel Garber, Magdalena J Koziol, Daniela Kenzelmann-Broz, Ahmad M Khalil, Or Zuk, Ido Amit, Michal Rabani, et al. A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 142 (3):409–419, 2010. doi:10.1016/j.cell.2010.06.040.

[109] Ana C Marques and Chris P Ponting. Intergenic lncRNAs and the evolution of gene expression. *Current opinion in genetics & development*, 27:48–53, 2014. doi:10.1016/j.gde.2014.03.009.

[110] Aurélie Kapusta, Zev Kronenberg, Vincent J Lynch, Xiaoyu Zhuo, LeeAnn Ramsay, Guillaume Bourque, Mark Yandell, and Cedric Feschotte. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet*, 9(4): e1003470, 2013. doi:10.1371/journal.pgen.1003470.

[111] Alex C Tuck, Kedar Nath Natarajan, Greggory M Rice, Jason Borawski, Fabio Mohn, Aneliya Rankova, Matyas Flemr, Alice Wenger, Razvan Nutiu, Sarah Teichmann, et al. Distinctive features of lincRNA gene expression suggest widespread rna-independent functions. *Life science alliance*, 1(4), 2018. doi:10.26508/lsa.201800124.

[112] Shea J Andrews and Joseph A Rothnagel. Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics*, 15(3):193–204, 2014. doi:10.1038/nrg3520.

[113] Zhe Ji, Ruisheng Song, Aviv Regev, and Kevin Struhl. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife*, 4:e08890, 2015. doi:10.7554/eLife.08890.

[114] Sebastiaan van Heesch, Maarten van Iterson, Jetse Jacobi, Sander Boymans, Paul B Essers, Ewart de Bruijn, Wensi Hao, Alyson W MacInnes, Edwin Cuppen, and Marieke Simonis. Extensive localization of long noncoding RNAs to the cytosol and mono-and polyribosomal complexes. *Genome biology*, 15(1):1–12, 2014. doi:10.1186/gb-2014-15-1-r6.

[115] Margarita Schlackow, Takayuki Nojima, Tomas Gomes, Ashish Dhir, Maria Carmo-Fonseca, and Nick J Proudfoot. Distinctive patterns of transcription and rna processing for human lincRNAs. *Molecular cell*, 65(1):25–38, 2017. doi:10.1016/j.molcel.2016.11.029.

[116] Anamaria Necsulea, Magali Soumillon, Maria Warnefors, Angélica Liechti, Tasman Daish, Ulrich Zeller, Julie C Baker, Frank Grützner, and Henrik Kaessmann. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, 505(7485):635–640, 2014. doi:10.1038/nature12943.

[117] Sara Massone, Eleonora Ciarlo, Serena Vella, Mario Nizzari, Tullio Florio, Claudio Russo, Ranieri Cancedda, and Aldo Pagano. NDM29, a RNA polymerase III-dependent non coding RNA, promotes amyloidogenic processing of APP and amyloid $\beta$ secretion. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1823(7):1170–1177, 2012. doi:10.1016/j.bbamcr.2012.05.001.

[118] Federico Ariel, Natali Romero-Barrios, Teddy Jégu, Moussa Benhamed, and Martin Crespi. Battles and hijacks: noncoding transcription in plants. *Trends in plant science*, 20(6):362–371, 2015. doi:10.1016/j.tplants.2015.03.003.

[119] Tanvir Alam, Yulia A Medvedeva, Hui Jia, James B Brown, Leonard Lipovich, and Vladimir B Bajic. Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes. *PloS one*, 9(10):e109443, 2014. doi:10.1371/journal.pone.0109443.

[120] Ashish Dhir, Somdutta Dhir, Nick J Proudfoot, and Catherine L Jopling. Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. *Nature structural & molecular biology*, 22(4):319, 2015. doi:10.1038/nsmb.2982.

[121] Moran N Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L Rinn. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development*, 25(18):1915–1927, 2011. doi:10.1101/gad.17446611.

[122] Aurélie Kapusta and Cédric Feschotte. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends in Genetics*, 30(10):439–452, 2014. doi:10.1016/j.tig.2014.08.004.

[123] Ana C Ayupe, Ana C Tahira, Lauren Camargo, Felipe C Beckedorff, Sergio Verjovski-Almeida, and Eduardo M Reis. Global analysis of biogenesis, stability and sub-cellular localization of lncRNAs mapping to intragenic regions of the human genome. *RNA biology*, 12(8):877–892, 2015. doi:10.1080/15476286.2015.1062960.

[124] Ling-Ling Chen. Linking long noncoding RNA localization and function. *Trends in biochemical sciences*, 41(9):761–772, 2016. doi:10.1016/j.tibs.2016.07.003.

[125] Ezgi Hacisuleyman, Loyal A Goff, Cole Trapnell, Adam Williams, Jorge Henao-Mejia, Lei Sun, Patrick McClanahan, David G Hendrickson, Martin Sauvageau, David R Kelley, et al. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nature structural & molecular biology*, 21(2):198, 2014. doi:10.1038/nsmb.2764.

[126] Bing Zhang, Lalith Gunawardane, Farshad Niazi, Fereshteh Jahanbani, Xin Chen, and Saba Valadkhan. A novel RNA motif mediates the strict nuclear localization of a long noncoding RNA. *Molecular and cellular biology*, 34(12):2318–2329, 2014. doi:10.1128/mcb.01673-13.

[127] Dan Grandér and Per Johnsson. Pseudogene-expressed RNAs: emerging roles in gene regulation and disease. In *Long Non-coding RNAs in Human Disease*, pages 111–126. Springer, 2015. doi:10.1007/82_2015_442.

[128] Ivelisse Cajigas, David E Leib, Jesse Cochrane, Hao Luo, Kelsey R Swyter, Sean Chen, Brian S Clark, James Thompson, John R Yates, Robert E Kingston, et al. Evf2 lncRNA/BRG1/DLX1 interactions reveal rna-dependent inhibition of chromatin remodeling. *Development*, 142(15): 2641–2652, 2015. doi:10.1242/dev.126318.

[129] Jessica Greenwood and Julia Promisel Cooper. Non-coding telomeric and subtelomeric transcripts are differentially regulated by telomeric and heterochromatin assembly factors in fission yeast. *Nucleic acids research*, 40(7):2956–2963, 2012. doi:10.1093/nar/gkr1155.

[130] Wenbo Li, Dimple Notani, Qi Ma, Bogdan Tanasa, Esperanza Nunez, Aaron Yun Chen, Daria Merkurjev, Jie Zhang, Kenneth Ohgi, Xiaoyuan Song, et al. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature*, 498(7455):516–520, 2013. doi:10.1038/nature12210.

[131] Evgenia Ntini, Aino I Järvelin, Jette Bornholdt, Yun Chen, Mette Boyd, Mette Jørgensen, Robin Andersson, Ilka Hoof, Aleks Schein, Peter R Andersen, et al. Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nature structural & molecular biology*, 20(8):923, 2013. doi:10.1038/nsmb.2640.

[132] Tim R Mercer and John S Mattick. Structure and function of long noncoding RNAs in epigenetic regulation. *Nature structural & molecular biology*, 20(3):300–307, 2013. doi:10.1038/nsmb.2480.

[133] Jeffrey J Quinn and Howard Y Chang. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, 17(1):47, 2016. doi:10.1038/nrg.2015.10.

[134] Pei Han and Ching-Pin Chang. Long non-coding RNA and chromatin remodeling. *RNA biology*, 12(10):1094–1098, 2015. doi:10.1080/15476286.2015.1063770.

[135] Je-Hyun Yoon, Kotb Abdelmohsen, Jiyoung Kim, Xiaoling Yang, Jennifer L Martindale, Kumiko Tominaga-Yamanaka, Elizabeth J White, Arturo V Orjalo, John L Rinn, Stefan G Kreft, et al. Scaffold function of long non-coding RNA hotair in protein ubiquitination. *Nature communications*, 4(1):1–14, 2013. doi:10.1038/ncomms3939.

[136] Tomohiro Yamazaki and Tetsuro Hirose. The building process of the functional paraspeckle with long non-coding RNAs. *Front Biosci (Elite Ed)*, 7(1):1–41, 2015. doi:10.2741/E715.

[137] Pushkar Malakar, Asaf Shilo, Adi Mogilevsky, Ilan Stein, Eli Pikarsky, Yuval Nevo, Hadar Benyamini, Sharona Elgavish, Xinying Zong, Kannanganattu V Prasanth, et al. Long noncoding RNA MALAT1 promotes hepatocellular carcinoma development by SRSF1 upregulation and mTOR activation. *Cancer research*, 77(5):1155–1167, 2017. doi:10.1158/0008-5472.CAN-16-1508.

[138] Tanmoy Mondal, Santhilal Subhash, Roshan Vaid, Stefan Enroth, Sireesha Uday, Björn Reinius, Sanhita Mitra, Arif Mohammed, Alva Rani James, Emily Hoberg, et al. MEG3 long noncoding RNA regulates the TGF-$\beta$ pathway genes through formation of RNA–DNA triplex structures. *Nature communications*, 6:7743, 2015. doi:10.1038/ncomms8743.

[139] Anton Scott Goustin, Pattaraporn Thepsuwan, Mary Ann Kosir, and Leonard Lipovich. The growth-arrest-specific (gas)-5 long non-coding RNA: a fascinating lncRNA widely expressed in cancers. *Non-coding RNA*, 5(3):46, 2019. doi:10.3390/ncrna5030046.

[140] Yvonne Tay, John Rinn, and Pier Paolo Pandolfi. The multilayered complexity of ceRNA crosstalk and competition. *Nature*, 505(7483):344–352, 2014. doi:10.1038/nature12986.

[141] Rituparno Sen, Suman Ghosal, Shaoli Das, Subrata Balti, and Jayprokas Chakrabarti. Competing endogenous RNA: the key to posttranscriptional regulation. *The Scientific World Journal*, 2014, 2014. doi:10.1155/2014/896206.

[142] Xuefei Shi, Ming Sun, Hongbing Liu, Yanwen Yao, and Yong Song. Long non-coding RNAs: a new frontier in the study of human diseases. *Cancer letters*, 339(2):159–166, 2013. doi:10.1016/j.canlet.2013.06.013.

[143] Bijan K Dey, Karl Pfeifer, and Anindya Dutta. The h19 long noncoding RNA gives rise to microRNAs miR-675-3p and miR-675-5p to promote skeletal muscle differentiation and regeneration. *Genes & development*, 28(5):491–501, 2014. doi:10.1101/gad.234419.113.

[144] Raquel Boque-Sastre, Marta Soler, Cristina Oliveira-Mateos, Anna Portela, Catia Moutinho, Sergi Sayols, Alberto Villanueva, Manel Esteller, and Sonia Guil. Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proceedings of the National Academy of Sciences*, 112(18):5785–5790, 2015. doi:10.1073/pnas.1421197112.

[145] Jesse M Engreitz, Jenna E Haines, Elizabeth M Perez, Glen Munson, Jenny Chen, Michael Kane, Patrick E McDonel, Mitchell Guttman, and Eric S Lander. Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature*, 539(7629):452–455, 2016. doi:10.1038/nature20149.

[146] Chenguang Gong and Lynne E Maquat. lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature*, 470(7333):284–288, 2011. doi:10.1038/nature09701.

[147] Sungyul Lee, Florian Kopp, Tsung-Cheng Chang, Anupama Sataluri, Beibei Chen, Sushama Sivakumar, Hongtao Yu, Yang Xie, and Joshua T Mendell. Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell*, 164(1-2):69–80, 2016. doi:10.1016/j.cell.2015.12.017.

[148] Ailone Tichon, Noa Gil, Yoav Lubelsky, Tal Havkin Solomon, Doron Lemze, Shalev Itzkovitz, Noam Stern-Ginossar, and Igor Ulitsky. A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nature communications*, 7(1):1–10, 2016. doi:10.1038/ncomms12209.

[149] Teddy Jégu, Eric Aeby, and Jeannie T Lee. The X chromosome in space. *Nature Reviews Genetics*, 18(6):377, 2017. doi:10.1038/nrg.2017.17.

[150] KM Creamer and Jeanne B Lawrence. XIST RNA: a window into the broader role of RNA in nuclear chromosome architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1733):20160360, 2017. doi:10.1098/rstb.2016.0360.

[151] Christine Moulton Clemson, John A McNeil, Huntington F Willard, and Jeanne Bentley Lawrence. XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *The Journal of cell biology*, 132(3):259–275, 1996. doi:10.1083/jcb.132.3.259.

[152] Chun-Kan Chen, Mario Blanco, Constanza Jackson, Erik Aznauryan, Noah Ollikainen, Christine Surka, Amy Chow, Andrea Cerase, Patrick McDonel, and Mitchell Guttman. Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science*, 354(6311): 468–472, 2016. doi:10.1126/science.aae0047.

[153] Benoit Moindrot, Andrea Cerase, Heather Coker, Osamu Masui, Anne Grijzenhout, Greta Pintacuda, Lothar Schermelleh, Tatyana B Nesterova, and Neil Brockdorff. A pooled shRNA screen identifies Rbm15, Spen, and Wtap as factors required for Xist RNA-mediated silencing. *Cell reports*, 12(4):562–572, 2015. doi:10.1016/j.celrep.2015.06.053.

[154] Asun Monfort, Giulio Di Minin, Andreas Postlmayr, Remo Freimann, Fabiana Arieti, Stéphane Thore, and Anton Wutz. Identification of Spen as a crucial factor for Xist function through forward genetic screening in haploid embryonic stem cells. *Cell reports*, 12(4):554–561, 2015. doi:10.1016/j.celrep.2015.06.067.

[155] Patrick McDonel and Mitchell Guttman. Approaches for understanding the mechanisms of long noncoding RNA regulation of gene expression. *Cold Spring Harbor Perspectives in Biology*, 11 (12):a032151, 2019. doi:10.1101/cshperspect.a032151.

[156] Manuela Portoso, Roberta Ragazzini, Živa Brenčič, Arianna Moiani, Audrey Michaud, Ivaylo Vassilev, Michel Wassef, Nicolas Servant, Bruno Sargueil, and Raphaël Margueron. PRC 2 is dispensable for HOTAIR-mediated transcriptional repression. *The EMBO journal*, 36(8):981–994, 2017. doi:10.15252/embj.201695335.

[157] F. Kopp and J. T. Mendell. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell*, 172(3):393–407, 01 2018. doi:10.1016/j.cell.2018.01.011.

[158] Abigail F Groff, Diana B Sanchez-Gomez, Marcela ML Soruco, Chiara Gerhardinger, A Rasim Barutcu, Eric Li, Lara Elcavage, Olivia Plana, Lluvia V Sanchez, James C Lee, et al. In vivo characterization of Linc-p21 reveals functional cis-regulatory DNA elements. *Cell reports*, 16(8): 2178–2186, 2016. doi:10.1016/j.celrep.2016.07.050.

[159] Bethany Signal, Brian S Gloss, and Marcel E Dinger. Computational approaches for functional prediction and characterisation of long noncoding RNAs. *Trends in Genetics*, 32(10):620–637, 2016. doi:10.1016/j.tig.2016.08.004.

[160] Stefan Washietl, Manolis Kellis, and Manuel Garber. Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome research*, 24(4):616–628, 2014. doi:10.1101/gr.165035.113.

[161] Jasmina Ponjavic, Chris P Ponting, and Gerton Lunter. Functionality or transcriptional noise? evidence for selection within long noncoding RNAs. *Genome research*, 17(5):556–565, 2007. doi:10.1101/gr.6036807.

[162] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA package 2.0. *Algorithms for molecular biology*, 6(1):26, 2011. doi:10.1186/1748-7188-6-26.

[163] KA_RST Hoogsteen. The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. *Acta Crystallographica*, 16(9):907–916, 1963. doi:10.1107/s0365110x63002437.

[164] H. Zhang, C. Zhang, Z. Li, C. Li, X. Wei, B. Zhang, and Y. Liu. A New Method of RNA Secondary Structure Prediction Based on Convolutional Neural Network and Dynamic Programming. *Front Genet*, 10:467, 2019. doi:10.3389/fgene.2019.00467.

[165] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, Jan 1981. doi:10.1093/nar/9.1.133.

[166] C. Pegueroles and T. Gabaldón. Secondary structure impacts patterns of selection in human lncRNAs. *BMC Biol*, 14:60, 07 2016. doi:10.1186/s12915-016-0283-0.

[167] A. Zampetaki, A. Albrecht, and K. Steinhofel. Long Non-coding RNA Structure and Function: Is There a Link? *Front Physiol*, 9:1201, 2018. doi:10.3389/fphys.2018.01201.

[168] S. Busan and K. M. Weeks. Visualization of RNA structure models within the Integrative Genomics Viewer. *RNA*, 23(7):1012–1018, 07 2017. doi:10.1261/rna.060194.116.

[169] M. J. Smola, T. W. Christy, K. Inoue, C. O. Nicholson, M. Friedersdorf, J. D. Keene, D. M. Lee, J. M. Calabrese, and K. M. Weeks. SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proc Natl Acad Sci U S A*, 113(37):10322–10327, 09 2016. doi:10.1073/pnas.1600008113.

[170] Alisha N Jones and Michael Sattler. Challenges and perspectives for structural biology of lncRNAs—the example of the Xist lncRNA A-repeats. *Journal of Molecular Cell Biology*, 11 (10):845–859, 2019. doi:10.1093/jmcb/mjz086.

[171] I. V. Novikova, S. P. Hennelly, and K. Y. Sanbonmatsu. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic acids research*, 40(11):5034–5051, Jun 2012. doi:10.1093/nar/gks071.

[172] L. Gao, Y. Liu, S. Guo, R. Yao, L. Wu, L. Xiao, Z. Wang, Y. Liu, and Y. Zhang. Circulating Long Noncoding RNA HOTAIR is an Essential Mediator of Acute Myocardial Infarction. *Cell Physiol Biochem*, 44(4):1497–1508, 2017. doi:10.1159/000485588.

[173] Z. Xue, S. Hennelly, B. Doyle, A. A. Gulati, I. V. Novikova, K. Y. Sanbonmatsu, and L. A. Boyer. A G-Rich Motif in the lncRNA Braveheart Interacts with a Zinc-Finger Transcription Factor to Specify the Cardiovascular Lineage. *Mol Cell*, 64(1):37–50, 10 2016. doi:10.1016/j.molcel.2016.08.010.

[174] J. Singh, J. Hanson, K. Paliwal, and Y. Zhou. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat Commun*, 10(1): 5407, 11 2019. doi:10.1038/s41467-019-13395-9.

[175] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA*, 18(5):900–914, May 2012. doi:10.1261/rna.029041.111.

[176] X. Zhang, R. Hong, W. Chen, M. Xu, and L. Wang. The role of long noncoding RNA in major human disease. *Bioorg Chem*, 92:103214, 11 2019. doi:10.1016/j.bioorg.2019.103214.

[177] Luka Bolha, Metka Ravnik-Glavač, and Damjan Glavač. Long noncoding RNAs as biomarkers in cancer. *Disease markers*, 2017, 2017. doi:10.1155/2017/7243968.

[178] M. Luo, Z. Li, W. Wang, Y. Zeng, Z. Liu, and J. Qiu. Long non-coding RNA H19 increases bladder cancer metastasis by associating with EZH2 and inhibiting E-cadherin expression. *Cancer Lett*, 333(2):213–221, Jun 2013. doi:10.1016/j.juro.2013.07.063.

[179] X. Zhou, C. Yin, Y. Dang, F. Ye, and G. Zhang. Identification of the long non-coding RNA H19 in plasma as a novel biomarker for diagnosis of gastric cancer. *Sci Rep*, 5:11516, Jun 2015. doi:10.1038/srep11516.

[180] D. Tan, Y. Wu, L. Hu, P. He, G. Xiong, Y. Bai, and K. Yang. Long noncoding RNA H19 is up-regulated in esophageal squamous cell carcinoma and promotes cell proliferation and metastasis. *Dis Esophagus*, 30(1):1–9, Jan 2017. doi:10.1111/dote.12481.

[181] W. Yang, N. Ning, and X. Jin. The lncRNA H19 Promotes Cell Proliferation by Competitively Binding to miR-200a and Derepressing $\beta$-Catenin Expression in Colorectal Cancer. *Biomed Res Int*, 2017:2767484, 2017. doi:10.1155/2017/2767484.

[182] M. H. Yang, Z. Y. Hu, C. Xu, L. Y. Xie, X. Y. Wang, S. Y. Chen, and Z. G. Li. MALAT1 promotes colorectal cancer cell proliferation/migration/invasion via PRKA kinase anchor protein 9. *Biochim Biophys Acta*, 1852(1):166–174, Jan 2015. doi:10.1016/j.bbadis.2014.11.013.

[183] Y. Shi, Y. Wang, W. Luan, P. Wang, T. Tao, J. Zhang, J. Qian, N. Liu, and Y. You. Long non-coding RNA H19 promotes glioma cell invasion by deriving miR-675. *PLoS One*, 9(1):e86295, 2014. doi:10.1371/journal.pone.0086295.

[184] T. Gutschner, M. Hämmerle, M. Eissmann, J. Hsu, Y. Kim, G. Hung, A. Revenko, G. Arun, M. Stentrup, M. Gross, M. Zörnig, A. R. MacLeod, D. L. Spector, and S. Diederichs. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res*, 73(3):1180–1189, Feb 2013. doi:10.1158/0008-5472.CAN-12-2850.

[185] S. Ren, F. Wang, J. Shen, Y. Sun, W. Xu, J. Lu, M. Wei, C. Xu, C. Wu, Z. Zhang, X. Gao, Z. Liu, J. Hou, J. Huang, and Y. Sun. Long non-coding RNA metastasis associated in lung adenocarcinoma transcript 1 derived miniRNA as a novel plasma-based biomarker for diagnosing prostate cancer. *Eur J Cancer*, 49(13):2949–2959, Sep 2013. doi:10.1016/j.ejca.2013.04.026.

[186] R. Merola, L. Tomao, A. Antenucci, I. Sperduti, S. Sentinelli, S. Masi, C. Mandoj, G. Orlandi, R. Papalia, S. Guaglianone, M. Costantini, G. Cusumano, G. Cigliana, P. Ascenzi, M. Gallucci, and L. Conti. PCA3 in prostate cancer and tumor aggressiveness detection on 407 high-risk patients: a National Cancer Institute experience. *J Exp Clin Cancer Res*, 34:15, Feb 2015. doi:10.1186/s13046-015-0127-8.

[187] F. Q. Nie, M. Sun, J. S. Yang, M. Xie, T. P. Xu, R. Xia, Y. W. Liu, X. H. Liu, E. B. Zhang, K. H. Lu, and Y. Q. Shu. Long noncoding RNA ANRIL promotes non-small cell lung cancer cell proliferation and inhibits apoptosis by silencing KLF2 and P21 expression. *Mol Cancer Ther*, 14 (1):268–277, Jan 2015. doi:10.1158/1535-7163.MCT-14-0492.

[188] J. R. Prensner, M. K. Iyer, A. Sahu, I. A. Asangani, Q. Cao, L. Patel, I. A. Vergara, E. Davicioni, N. Erho, M. Ghadessi, R. B. Jenkins, T. J. Triche, R. Malik, R. Bedenis, N. McGregor, T. Ma, W. Chen, S. Han, X. Jing, X. Cao, X. Wang, B. Chandler, W. Yan, J. Siddiqui, L. P. Kunju, S. M. Dhanasekaran, K. J. Pienta, F. Y. Feng, and A. M. Chinnaiyan. The long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat Genet*, 45(11):1392–1398, Nov 2013. doi:10.1038/ng.2771.

[189] H. Yu, A. Xu, B. Wu, M. Wang, and Z. Chen. Long noncoding RNA NEAT1 promotes progression of glioma as a ceRNA by sponging miR-185-5p to stimulate DNMT1/mTOR signaling. *J Cell Physiol*, Aug 2020. doi:10.1002/jcp.29644.

[190] S. N. Fotuhi, M. Khalaj-Kondori, M. A. Hoseinpour Feizi, and M. Talebi. Long Non-coding RNA BACE1-AS May Serve as an Alzheimer's Disease Blood-Based Biomarker. *J Mol Neurosci*, 69(3):351–359, Nov 2019. doi:10.1007/s12031-019-01364-2.

[191] F. Xu, L. Jin, Y. Jin, Z. Nie, and H. Zheng. Long noncoding RNAs in autoimmune diseases. *J Biomed Mater Res A*, 107(2):468–475, 02 2019. doi:10.1002/jbm.a.36562.

[192] X. Li, H. Wang, B. Yao, W. Xu, J. Chen, and X. Zhou. lncRNA H19/miR-675 axis regulates cardiomyocyte apoptosis by targeting VDAC1 in diabetic cardiomyopathy. *Sci Rep*, 6:36340, 10 2016. doi:10.1038/srep36340.

[193] Y. Lyu, L. Bai, and C. Qin. Long noncoding RNAs in neurodevelopment and Parkinson's disease. *Animal Model Exp Med*, 2(4):239–251, Dec 2019. doi:10.1002/ame2.12093.

[194] J. M. Lorenzen and T. Thum. Long noncoding RNAs in kidney and cardiovascular diseases. *Nat Rev Nephrol*, 12(6):360–373, 06 2016. doi:10.1038/nrneph.2016.51.

[195] Zhenyu Bao, Zhen Yang, Zhou Huang, Yiran Zhou, Qinghua Cui, and Dong Dong. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic acids research*, 47(D1):D1034–D1037, 10 2018. ISSN 0305-1048. doi:10.1093/nar/gky905. URL `https://doi.org/10.1093/nar/gky905`.

[196] Alan M Turing. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer, 2009.

[197] Samantha Hayman. The Mcculloch-Pitts model. In *IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339)*, volume 6, pages 4438–4439. IEEE, 1999.

[198] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi:10.1038/323533a0.

[199] Feng-hsiung Hsu, Thomas Anantharaman, Murray Campbell, and Andreas Nowatzyk. A grandmaster chess machine. *Scientific American*, 263(4):44–51, 1990. doi:10.1038/scientificamerican1090-44.

[200] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[201] Daniel SW Ting, Yong Liu, Philippe Burlina, Xinxing Xu, Neil M Bressler, and Tien Y Wong. AI for medical imaging goes deep. *Nature medicine*, 24(5):539–540, 2018. doi:10.1038/s41591-018-0029-3.

[202] Tim Dettmers. Deep learning in a nutshell: Core concepts. *NVIDIA Devblogs*, 2015.

[203] Donald Michie, David J Spiegelhalter, CC Taylor, et al. Machine learning. *Neural and Statistical Classification*, 13(1994):1–298, 1994.

[204] Tom Michael Mitchell. *The discipline of machine learning*, volume 9. Carnegie Mellon University, School of Computer Science, Machine Learning, 2006.

[205] A. Saxe, S. Nelli, and C. Summerfield. If deep learning is the answer, what is the question? *Nat Rev Neurosci*, Nov 2020. doi:10.1038/s41583-020-00395-8.

[206] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi:10.1038/nature14539.

[207] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[208] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[209] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3626–3633, 2013. doi:10.1109/CVPR.2013.465.

[210] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[211] Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013.

[212] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

[213] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv*, pages arXiv–1909, 2019.

[214] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. The MIT Press Cambridge, 2016.

[215] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. DeepFace: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. doi:10.1109/CVPR.2014.220.

[216] Ian Goodfellow, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Multi-prediction deep Boltzmann machines. In *Advances in Neural Information Processing Systems*, pages 548–556, 2013.

[217] Douglas C Montgomery, G Geoffrey Vining, and Elizabeth A Peck. *Introduction to linear regression analysis*. Wiley, 2012.

[218] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105, 2004.

[219] Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated Machine Learning*, pages 3–33. Springer, Cham, 2019. doi:10.1007/978-3-030-05318-5_1.

[220] S Sra, S Nowozin, and SJ Wright. *Optimization for Machine Learning*. MIT Press, 2011. doi:10.7551/mitpress/8996.001.0001.

[221] Harald Cramér. *Mathematical methods of statistics*, volume 43. Princeton university press, 1999.

[222] C Radhakrishna Rao. Information and the accuracy attainable in the estimation of statistical parameters. *Reson. J. Sci. Educ*, 20:78–90, 1945.

[223] Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995. doi:10.1162/neco.1995.7.2.219.

[224] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010. doi:10.1007/978-3-7908-2604-3-16.

[225] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936. doi:10.1111/j.1469-1809.1936.tb02137.x.

[226] Horace B Barlow. Unsupervised learning. *Neural computation*, 1(3):295–311, 1989.

[227] Geoffrey E Hinton, Peter Dayan, and Michael Revow. Modeling the manifolds of images of handwritten digits. *IEEE transactions on Neural Networks*, 8(1):65–74, 1997. doi:10.1109/72.554192.

[228] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998. doi:10.1109/5254.708428.

[229] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011. doi:10.1145/1961189.1961199.

[230] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995. doi:10.1109/ICDAR.1995.598994.

[231] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. doi:10.1023/A:1010933404324.

[232] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. doi:10.1007/bf00058655.

[233] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. doi:10.1098/rsta.2015.0202.

[234] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. Constrained k-means clustering with background knowledge. In *Icml*, volume 1, pages 577–584, 2001.

[235] John A Hartigan and Manchek A Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979. doi:10.2307/2346830.

[236] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. TensorFlow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283, 2016.

[237] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[238] Régis Vaillant, Christophe Monrocq, and Yann Le Cun. Original approach for the localisation of objects in images. *IEE Proceedings-Vision, Image and Signal Processing*, 141(4):245–250, 1994.

[239] Dan CireşAn, Ueli Meier, Jonathan Masci, and Jürgen Schmidhuber. Multi-column deep neural network for traffic sign classification. *Neural networks*, 32:333–338, 2012.

[240] Feng Ning, Damien Delhomme, Yann LeCun, Fabio Piano, Léon Bottou, and Paolo Emilio Barbano. Toward automatic phenotyping of developing embryos from videos. *IEEE Transactions on Image Processing*, 14(9):1360–1371, 2005. doi:10.1109/TIP.2005.852470.

[241] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[242] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[243] Michael Buckland and Fredric Gey. The relationship between precision and recall. *Journal of the American Society for Information Science*, 45(1):17, 1994. doi:10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L.

[244] Warren J Ewens and Gregory R Grant. *Statistical methods in bioinformatics: an introduction*. Springer Science & Business Media, 2006.

[245] David Hand and Peter Christen. A note on using the F-measure for evaluating record linkage algorithms. *Statistics and Computing*, 28(3):539–547, 2018. doi:10.1007/s11222-017-9746-6.

[246] Nitesh V Chawla. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer, 2009. doi:10.1007/978-0-387-09823-4_45.

[247] Andrew P Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997. doi:10.1016/S0031-3203(96)00142-2.

[248] Brian W Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975. doi:10.1016/0005-2795(75)90109-9.

[249] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020. doi:10.1186/s12864-019-6413-7.

[250] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960. doi:10.1177/001316446002000104.

[251] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276, 2012. doi:10.11613/bm.2012.031.

[252] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1): 193–218, 1985. doi:10.1007/BF01908075.

[253] John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. The MIT Press, 2015. ISBN 0262029448.

[254] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[255] John A Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988. doi:10.1126/science.3287615.

[256] Nathalie Japkowicz. Concept-learning in the presence of between-class and within-class imbalances. In *Conference of the Canadian society for computational studies of intelligence*, pages 67–77. Springer, 2001. doi:10.1007/3-540-45153-6_7.

[257] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In *Icml*, volume 97, pages 179–186, 1997.

[258] J. Kim, D. Tae, and J. Seok. A survey of missing data imputation using generative adversarial networks. In *2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 454–456, 2020. doi:10.1109/ICAIIC48513.2020.9065044.

[259] Michael B Clark, Paulo P Amaral, Felix J Schlesinger, Marcel E Dinger, Ryan J Taft, John L Rinn, Chris P Ponting, Peter F Stadler, Kevin V Morris, Antonin Morillon, et al. The reality of pervasive transcription. *PLoS Biol*, 9(7):e1000625, 2011. doi:10.1371/journal.pbio.1000625.

[260] Sarah E Hunt, William McLaren, Laurent Gil, Anja Thormann, Helen Schuilenburg, Dan Sheppard, Andrew Parton, Irina M Armean, Stephen J Trevanion, Paul Flicek, et al. Ensembl variation resources. *Database*, 2018, 2018. doi:10.1093/database/bay119.

[261] David Mas-Ponte, Joana Carlevaro-Fita, Emilio Palumbo, Toni Hermoso Pulido, Roderic Guigo, and Rory Johnson. LncATLAS database for subcellular localization of long noncoding RNAs. *RNA*, 23(7):1080–1087, 2017. doi:10.1261/rna.060814.117.

[262] Lesca M Holdt, Steve Hoffmann, Kristina Sass, David Langenberger, Markus Scholz, Knut Krohn, Knut Finstermeier, Anika Stahringer, Wolfgang Wilfert, Frank Beutner, et al. Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS Genet*, 9(7):e1003588, 2013. doi:10.1371/journal.pgen.1003588.

[263] Esra Bozgeyik, Yusuf Ziya Igci, Mevan F Sami Jacksi, Kaifee Arman, Serdar A Gurses, Ibrahim Bozgeyik, Elif Pala, Onder Yumrutas, Ebru Temiz, and Mehri Igci. A novel variable exonic region and differential expression of LINC00663 non-coding RNA in various cancer cell lines and normal human tissue samples. *Tumor Biology*, 37(7):8791–8798, 2016. doi:10.1007/s13277-015-4782-3.

[264] Tim R Mercer, Dagmar Wilhelm, Marcel E Dinger, Giulia Solda, Darren J Korbie, Evgeny A Glazov, Vy Truong, Maren Schwenke, Cas Simons, Klaus I Matthaei, et al. Expression of distinct RNAs from 3' untranslated regions. *Nucleic acids research*, 39(6):2393–2403, 2011. doi:10.1093/nar/gkq1158.

[265] Jan Engelhardt and Peter F Stadler. Evolution of the unspliced transcriptome. *BMC evolutionary biology*, 15(1):166, 2015. doi:10.1186/s12862-015-0437-7.

[266] Jason G Cyster and Christopher DC Allen. B cell responses: cell interaction dynamics and decisions. *Cell*, 177(3):524–540, 2019. doi:10.1016/j.cell.2019.03.016.

[267] Shaoying Li, Ken H Young, and L Jeffrey Medeiros. Diffuse large B-cell lymphoma. *Pathology*, 50(1):74–87, 2018. doi:10.1016/j.pathol.2017.09.006.

[268] Antonino Carbone, Sandrine Roulland, Annunziata Gloghini, Anas Younes, Gottfried von Keudell, Armando López-Guillermo, and Jude Fitzgibbon. Follicular lymphoma. *Nature Reviews Disease Primers*, 5(1):1–20, 2019. doi:10.1038/s41572-019-0132-x.

[269] Chi Young Ok and L Jeffrey Medeiros. High-grade B-cell lymphoma: a term re-purposed in the revised WHO classification. *Pathology*, 52(1):68–77, 2020. doi:10.1016/j.pathol.2019.09.008.

[270] GENCODE releases. URL `https://www.gencodegenes.org/human/releases.html`. Accessed on March 20, 2016.

[271] Paul Flicek, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, et al. Ensembl 2014. *Nucleic acids research*, 42(D1):D749–D755, 2014. doi:10.1093/nar/gkt1196.

[272] Yi Zhao, Hui Li, Shuangsang Fang, Yue Kang, Wei Wu, Yajing Hao, Ziyang Li, Dechao Bu, Ninghui Sun, Michael Q Zhang, et al. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic acids research*, 44(D1):D203–D208, 2016. doi:10.1093/nar/gkv1252.

[273] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, Jörg Vogel, Peter F Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9):e1000502, 2009. doi:10.1371/journal.pcbi.1000502.

[274] Ensembl 83, . URL `ftp://ftp.ensembl.org/pub/release-60/gtf/`. Accessed on June 12, 2017.

[275] Ensembl 60, . URL `ftp://ftp.ensembl.org/pub/release-83/gtf/`. Accessed on March 9, 2016.

[276] Jonathan D Ellis, Miriam Barrios-Rodiles, Recep Çolak, Manuel Irimia, TaeHyung Kim, John A Calarco, Xinchen Wang, Qun Pan, Dave O'Hanlon, Philip M Kim, et al. Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular cell*, 46(6):884–892, 2012. doi:10.1016/j.molcel.2012.05.037.

[277] Yue Cao, Thomas Andrew Geddes, Jean Yee Hwa Yang, and Pengyi Yang. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*, pages 1–9, 2020. doi:10.1038/s42256-020-0217-y.

[278] Manel Esteller. Non-coding RNAs in human disease. *Nature reviews genetics*, 12(12):861–874, 2011. doi:10.1038/nrg3074.

[279] Noorul Amin, Annette McGrath, and Yi-Ping Phoebe Chen. Evaluation of deep learning in non-coding RNA classification. *Nature Machine Intelligence*, 1(5):246–256, 2019. doi:10.1038/s42256-019-0051-2.

[280] James W Fickett. Recognition of protein coding regions in DNA sequences. *Nucleic acids research*, 10(17):5303–5318, 1982. doi:10.1093/nar/10.17.5303.

[281] Liguo Wang, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, and Wei Li. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research*, 41(6):e74–e74, 2013. doi:10.1093/nar/gkt006.

[282] Siyu Han, Yanchun Liang, Ying Li, and Wei Du. Long noncoding RNA identification: comparing machine learning based tools for long noncoding transcripts discrimination. *BioMed research international*, 2016, 2016. doi:10.1155/2016/8496165.

[283] Lei Kong, Yong Zhang, Zhi-Qiang Ye, Xiao-Qiao Liu, Shu-Qi Zhao, Liping Wei, and Ge Gao. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research*, 35(suppl_2):W345–W349, 2007. doi:10.1093/nar/gkm391.

[284] James W Fickett and Chang-Shung Tung. Assessment of protein coding measures. *Nucleic acids research*, 20(24):6441–6450, 1992. doi:10.1093/nar/20.24.6441.

[285] Siyu Han, Yanchun Liang, Qin Ma, Yangyi Xu, Yu Zhang, Wei Du, Cankun Wang, and Ying Li. LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. *Briefings in bioinformatics*, 20(6):2009–2027, 2019. doi:10.1093/bib/bby065.

[286] Uberto Pozzoli, Giorgia Menozzi, Matteo Fumagalli, Matteo Cereda, Giacomo P Comi, Rachele Cagliani, Nereo Bresolin, and Manuela Sironi. Both selective and neutral processes drive GC content evolution in the human genome. *BMC evolutionary biology*, 8(1):99, 2008. doi:10.1186/1471-2148-8-99.

[287] Aimin Li, Junying Zhang, and Zhongyin Zhou. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC bioinformatics*, 15(1): 311, 2014. doi:10.1186/1471-2105-15-311.

[288] Jessime M Kirk, Susan O Kim, Kaoru Inoue, Matthew J Smola, David M Lee, Megan D Schertzer, Joshua S Wooten, Allison R Baker, Daniel Sprague, David W Collins, et al. Functional classification of long non-coding RNAs by k-mer content. *Nature genetics*, 50(10):1474–1482, 2018. doi:10.1038/s41588-018-0207-8.

[289] Michael F Lin, Irwin Jungreis, and Manolis Kellis. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–i282, 2011. doi:10.1093/bioinformatics/btr209.

[290] Martin C Frith, Timothy L Bailey, Takeya Kasukawa, Flavio Mignone, Sarah K Kummerfeld, Martin Madera, Sirisha Sunkara, Masaaki Furuno, Carol J Bult, John Quackenbush, et al. Discrimination of non-protein-coding transcripts from protein-coding mRNA. *RNA biology*, 3(1): 40–48, 2006. doi:10.4161/rna.3.1.2789.

[291] Jinfeng Liu, Julian Gough, and Burkhard Rost. Distinguishing protein-coding from non-coding RNAs through support vector machines. *PLoS Genet*, 2(4):e29, 2006. doi:10.1371/journal.pgen.0020029.

[292] G.S.C. Slater. *Algorithms for the analysis of ESTs*. PhD thesis, University of Cambridge, 1998.

[293] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997. doi:10.1093/nar/25.17.3389.

[294] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169, 2017. doi:10.1093/nar/gkw1099.

[295] Yu-Jian Kang, De-Chang Yang, Lei Kong, Mei Hou, Yu-Qi Meng, Liping Wei, and Ge Gao. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic acids research*, 45(W1):W12–W16, 2017. doi:10.1093/nar/gkx428.

[296] Liang Sun, Haitao Luo, Dechao Bu, Guoguang Zhao, Kuntao Yu, Changhai Zhang, Yuanning Liu, Runsheng Chen, and Yi Zhao. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic acids research*, 41(17):e166–e166, 2013. doi:10.1093/nar/gkt646.

[297] Jin-Cheng Guo, Shuang-Sang Fang, Yang Wu, Jian-Hua Zhang, Yang Chen, Jing Liu, Bo Wu, Jia-Rui Wu, En-Min Li, Li-Yan Xu, et al. CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. *Nucleic acids research*, 47(W1):W516–W522, 2019. doi:10.1093/nar/gkz400.

[298] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. doi:10.1145/2939672.2939785.

[299] Stephan H Bernhart, Ivo L Hofacker, and Peter F Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–615, 2006. doi:10.1093/bioinformatics/btk014.

[300] Achuthsankar S Nair and Sivarama Pillai Sreenadhan. A coding measure scheme employing electron-ion interaction pseudopotential (eiip). *Bioinformation*, 1(6):197, 2006.

[301] Kun Sun, Xiaona Chen, Peiyong Jiang, Xiaofeng Song, Huating Wang, and Hao Sun. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. *BMC genomics*, 14(S2):S7, 2013. doi:10.1186/1471-2164-14-S2-S7.

[302] Adam Siepel, Gill Bejerano, Jakob S Pedersen, Angie S Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W Hillier, Stephen Richards, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8): 1034–1050, 2005. doi:10.1101/gr.3715005.

[303] Adam Siepel and David Haussler. Phylogenetic hidden Markov models. In *Statistical methods in molecular evolution*, pages 325–351. Springer, 2005. doi:10.1007/0-387-27733-1_12.

[304] Lei Sun, Hui Liu, Lin Zhang, and Jia Meng. lncRScan-SVM: a tool for predicting long non-coding RNAs using support vector machine. *PloS one*, 10(10):e0139654, 2015. doi:10.1371/journal.pone.0139654.

[305] Valentin Wucher, Fabrice Legeai, Benoit Hedan, Guillaume Rizk, Lætitia Lagoutte, Tosso Leeb, Vidhya Jagannathan, Edouard Cadieu, Audrey David, Hannes Lohi, et al. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic acids research*, 45(8):e57–e57, 2017. doi:10.1093/nar/gkw1306.

[306] Rujira Achawanantakun, Jiao Chen, Yanni Sun, and Yuan Zhang. LncRNA-ID: Long non-coding RNA IDentification using balanced random forests. *Bioinformatics*, 31(24):3897–3905, 2015. doi:10.1093/bioinformatics/btv480.

[307] Marilyn Kozak. Initiation of translation in prokaryotes and eukaryotes. *Gene*, 234(2):187–208, 1999. doi:10.1016/S0378-1119(99)00210-3.

[308] Heng Xu, Ping Wang, Yujie Fu, Yufang Zheng, Quan Tang, Lizhen Si, Jin You, Zhenguo Zhang, Yufei Zhu, Li Zhou, et al. Length of the ORF, position of the first AUG and the Kozak motif are important factors in potential dual-coding transcripts. *Cell research*, 20(4):445–457, 2010. doi:10.1038/cr.2010.25.

[309] Cong Pian, Guangle Zhang, Zhi Chen, Yuanyuan Chen, Jin Zhang, Tao Yang, and Liangyun Zhang. LncRNApred: Classification of long non-coding RNAs and protein-coding transcripts by the ensemble algorithm with a new hybrid feature. *PloS one*, 11(5):e0154567, 2016. doi:10.1371/journal.pone.0154567.

[310] Junghwan Baek, Byunghan Lee, Sunyoung Kwon, and Sungroh Yoon. LncRNAnet: long non-coding RNA identification using deep learning. *Bioinformatics*, 34(22):3889–3897, 2018. doi:10.1093/bioinformatics/bty418.

[311] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi:10.1109/5.726791.

[312] Cheng Yang, Longshu Yang, Man Zhou, Haoling Xie, Chengjiu Zhang, May D Wang, and Huaiqiu Zhu. LncADeep: an ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics*, 34(22):3825–3834, 2018. doi:10.1093/bioinformatics/bty428.

[313] Yongchu Liu, Jiangtao Guo, Gangqing Hu, and Huaiqiu Zhu. Gene prediction in metagenomic fragments based on the SVM algorithm. In *BMC bioinformatics*, volume 14, page S12. Springer, 2013. doi:10.1186/1471-2105-14-S5-S12.

[314] Rashmi Tripathi, Sunil Patel, Vandana Kumari, Pavan Chakraborty, and Pritish Kumar Varadwaj. DeepLNC, a long non-coding RNA prediction tool using deep neural network. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 5(1):21, 2016. doi:10.1007/s13721-016-0129-2.

[315] Rolf Backofen, Stephan H Bernhart, Christoph Flamm, Claudia Fried, Guido Fritzsch, Jörg Hackermüller, Jana Hertel, Ivo L Hofacker, Kristin Missal, Axel Mosig, et al. RNAs everywhere: genome-wide annotation of structured RNAs. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 308(1):1–25, 2007. doi:10.1002/jez.b.21130.

[316] David P Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009. doi:10.1016/j.cell.2009.01.002.

[317] David P Bartel. Metazoan MicroRNAs. *Cell*, 173(1):20–51, 2018. doi:10.1016/j.cell.2018.03.006.

[318] Minju Ha and V Narry Kim. Regulation of microRNA biogenesis. *Nature reviews Molecular cell biology*, 15(8):509–524, 2014. doi:10.1038/nrm3838.

[319] Sarah Roush and Frank J Slack. The let-7 family of microRNAs. *Trends in cell biology*, 18(10): 505–516, 2008. doi:10.1016/j.tcb.2008.07.007.

[320] Alexander Sasse, Kaitlin U Laverty, Timothy R Hughes, and Quaid D Morris. Motif models for RNA-binding proteins. *Current Opinion in Structural Biology*, 53:115–123, 2018. doi:10.1016/j.sbi.2018.08.001.

[321] Giorgio Dieci, Milena Preti, and Barbara Montanini. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics*, 94(2):83–88, 2009. doi:10.1016/j.ygeno.2009.05.002.

[322] Joanna Kufel and Pawel Grzechnik. Small nucleolar RNAs tell a different tale. *Trends in Genetics*, 35(2):104–117, 2019. doi:10.1016/j.tig.2018.11.005.

[323] Søren Lykke-Andersen, Yun Chen, Britt R Ardal, Berit Lilje, Johannes Waage, Albin Sandelin, and Torben Heick Jensen. Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes. *Genes & development*, 28(22):2498–2517, 2014. doi:10.1101/gad.246538.114.

[324] Jinfei Tong, Xudong Ma, Hailan Yu, and Jianhua Yang. SNHG15: a promising cancer-related long noncoding RNA. *Cancer management and research*, 11:5961, 2019. doi:10.2147/CMAR.S208054.

[325] Wei Zhao, Xiaozhou Ma, Lina Liu, Qingyu Chen, Zihao Liu, Zheng Zhang, Shiqing Ma, Zhonghou Wang, Hongfa Li, Zuomin Wang, et al. SNHG20: A vital lncRNA in multiple human cancers. *Journal of cellular physiology*, 234(9):14519–14525, 2019. doi:10.1002/jcp.28143.

[326] Ziqiang Zhang, Z Zhu, K Watabe, X Zhang, C Bai, M Xu, F Wu, and YY Mo. Negative regulation of lncRNA GAS5 by miR-21. *Cell Death & Differentiation*, 20(11):1558–1568, 2013. doi:10.1038/cdd.2013.110.

[327] Ying Huang. The novel regulatory role of lnc RNA-miRNA-mRNA axis in cardiovascular diseases. *Journal of Cellular and Molecular Medicine*, 22(12):5768–5775, 2018. doi:10.1111/jcmm.13866.

[328] Yan Li, Zhenhui Zhao, Wei Liu, and Xun Li. SNHG3 functions as miRNA sponge to promote breast cancer cells growth through the metabolic reprogramming. *Applied Biochemistry and Biotechnology*, pages 1–16, 2020. doi:10.1007/s12010-020-03244-7.

[329] Huan Yang, Zheng Jiang, Shuang Wang, Yongbing Zhao, Xiaomei Song, Yufeng Xiao, and Shiming Yang. Long non-coding small nucleolar RNA host genes in digestive cancers. *Cancer medicine*, 8(18):7693–7704, 2019. doi:10.1002/cam4.2622.

[330] Marc P Hoeppner and Anthony M Poole. Comparative genomics of eukaryotic small nucleolar RNAs reveals deep evolutionary ancestry amidst ongoing intragenomic mobility. *BMC evolutionary biology*, 12(1):183, 2012. doi:10.1186/1471-2148-12-183.

[331] Qinyu Sun, Vidisha Tripathi, Je-Hyun Yoon, Deepak K Singh, Qinyu Hao, Kyung-Won Min, Sylvia Davila, Richard W Zealy, Xiao Ling Li, Maria Polycarpou-Schwarz, et al. MIR100 host gene-encoded lncRNAs regulate cell cycle by modulating the interaction between HuR and its target mRNAs. *Nucleic acids research*, 46(19):10405–10416, 2018. doi:10.1093/nar/gky696.

[332] Hsi-Yuan Huang, Yang-Chi-Dung Lin, Jing Li, Kai-Yao Huang, Sirjana Shrestha, Hsiao-Chin Hong, Yun Tang, Yi-Gang Chen, Chen-Nan Jin, Yuan Yu, et al. miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic acids research*, 48 (D1):D148–D154, 2020. doi:10.1093/nar/gkz896.

[333] Yongjie Sun, Xinghao Jia, Mingxing Wang, and Yiqi Deng. Long noncoding RNA MIR31HG abrogates the availability of tumor suppressor microRNA-361 for the growth of osteosarcoma. *Cancer Management and Research*, 11:8055, 2019. doi:10.2147/CMAR.S214569.

[334] Yong Tang, Xian Jin, Yin Xiang, Yu Chen, Cheng-xing Shen, Ya-chen Zhang, and Yi-gang Li. The lncRNA MALAT1 protects the endothelium against ox-LDL-induced dysfunction via upregulating the expression of the miR-22-3p target genes CXCR2 and AKT. *FEBS letters*, 589 (20):3189–3196, 2015. doi:10.1016/j.febslet.2015.08.046.

[335] Andrew Keniry, David Oxley, Paul Monnier, Michael Kyba, Luisa Dandolo, Guillaume Smits, and Wolf Reik. The H19 lincRNA is a developmental reservoir of miR-675 that suppresses growth and Igf1r. *Nature cell biology*, 14(7):659–665, 2012. doi:10.1038/ncb2521.

[336] Marc R Friedländer, Wei Chen, Catherine Adamidi, Jonas Maaskola, Ralf Einspanier, Signe Knespel, and Nikolaus Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology*, 26(4):407–415, 2008. doi:10.1038/nbt1394.

[337] Peng Jiang, Haonan Wu, Wenkai Wang, Wei Ma, Xiao Sun, and Zuhong Lu. miPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic acids research*, 35(suppl_2):W339–W344, 2007. doi:10.1093/nar/gkm368.

[338] Jana Hertel, Ivo L Hofacker, and Peter F Stadler. snoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24(2):158–164, 2008. doi:10.1093/bioinformatics/btm464.

[339] Qaisar Abbas, Syed Mansoor Raza, Azizuddin Ahmed Biyabani, and Muhammad Arfan Jaffar. A review of computational methods for finding non-coding RNA genes. *Genes*, 7(12):113, 2016. doi:10.3390/genes7120113.

[340] Georgios K Georgakilas, Andrea Grioni, Konstantinos G Liakos, Eliska Chalupova, Fotis C Plessas, and Panagiotis Alexiou. Multi-branch Convolutional Neural Network for Identification of Small Non-coding RNA genomic loci. *Scientific Reports*, 10(1):1–10, 2020. doi:10.1038/s41598-020-66454-3.

[341] Megha Ghildiyal and Phillip D Zamore. Small silencing RNAs: an expanding universe. *Nature Reviews Genetics*, 10(2):94–108, 2009. doi:10.1038/nrg2504.

[342] Rosalind C Lee, Rhonda L Feinbaum, and Victor Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993. doi:10.1016/0092-8674(93)90529-Y.

[343] Mariana Lagos-Quintana, Reinhard Rauhut, Winfried Lendeckel, and Thomas Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–858, 2001. doi:10.1126/science.1064921.

[344] Nelson C Lau, Lee P Lim, Earl G Weinstein, and David P Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–862, 2001. doi:10.1126/science.1065062.

[345] Jacek Krol, Inga Loedige, and Witold Filipowicz. The widespread regulation of microRNA biogenesis, function and decay. *Nature Reviews Genetics*, 11(9):597–610, 2010. doi:10.1038/nrg2843.

[346] Brandi N Davis-Dusenbery and Akiko Hata. Mechanisms of control of microRNA biogenesis. *The journal of biochemistry*, 148(4):381–392, 2010. doi:10.1093/jb/mvq096.

[347] Alex Mas Monteys, Ryan M Spengler, Ji Wan, Luis Tecedor, Kimberly A Lennox, Yi Xing, and Beverly L Davidson. Structure and activity of putative intronic miRNA promoters. *RNA*, 16(3): 495–505, 2010. doi:10.1261/rna.1731910.

[348] Tuan Anh Nguyen, Myung Hyun Jo, Yeon-Gil Choi, Joha Park, S Chul Kwon, Sungchul Hohng, V Narry Kim, and Jae-Sung Woo. Functional anatomy of the human microprocessor. *Cell*, 161 (6):1374–1387, 2015. doi:10.1016/j.cell.2015.05.010.

[349] Jinju Han, Yoontae Lee, Kyu-Hyeon Yeom, Jin-Wu Nam, Inha Heo, Je-Keun Rhee, Sun Young Sohn, Yunje Cho, Byoung-Tak Zhang, and V Narry Kim. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, 125(5):887–901, 2006. doi:10.1016/j.cell.2006.03.043.

[350] Wenwen Fang and David P Bartel. The menu of features that define primary microRNAs and enable de novo design of microRNA genes. *Molecular cell*, 60(1):131–145, 2015. doi:10.1016/j.molcel.2015.08.015.

[351] Vincent C Auyeung, Igor Ulitsky, Sean E McGeary, and David P Bartel. Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell*, 152 (4):844–858, 2013. doi:10.1016/j.cell.2013.01.031.

[352] Markus T Bohnsack, Kevin Czaplinski, and Dirk Görlich. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, 10(2):185–191, 2004.

[353] Elsebet Lund, Stephan Güttinger, Angelo Calado, James E Dahlberg, and Ulrike Kutay. Nuclear export of microRNA precursors. *Science*, 303(5654):95–98, 2004. doi:10.1126/science.1090599.

[354] György Hutvágner, Juanita McLachlan, Amy E Pasquinelli, Éva Bálint, Thomas Tuschl, and Phillip D Zamore. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293(5531):834–838, 2001. doi:10.1126/science.1062961.

[355] Haidi Zhang, Fabrice A Kolb, Lukasz Jaskiewicz, Eric Westhof, and Witold Filipowicz. Single processing center models for human Dicer and bacterial RNase III. *Cell*, 118(1):57–68, 2004. doi:10.1016/j.cell.2004.06.017.

[356] Shintaro Iwasaki, Maki Kobayashi, Mayuko Yoda, Yuriko Sakaguchi, Susumu Katsuma, Tsutomu Suzuki, and Yukihide Tomari. Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. *Molecular cell*, 39(2):292–299, 2010. doi:10.1016/j.molcel.2010.05.015.

[357] Hiroshi I Suzuki, Akihiro Katsura, Takahiko Yasuda, Toshihide Ueno, Hiroyuki Mano, Koichi Sugimoto, and Kohei Miyazono. Small-RNA asymmetry is directly driven by mammalian argonautes. *Nature structural & molecular biology*, 22(7):512–521, 2015. doi:10.1038/nsmb.3050.

[358] Eric Huntzinger and Elisa Izaurralde. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Reviews Genetics*, 12(2):99–110, 2011. doi:10.1038/nrg2936.

[359] Jacob O'Brien, Heyam Hayder, Yara Zayed, and Chun Peng. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in endocrinology*, 9:402, 2018. doi:10.3389/fendo.2018.00402.

[360] Christopher R Sibley, Yiqi Seow, Sheena Saayman, Krijn K Dijkstra, Samir El Andaloussi, Marc S Weinberg, and Matthew JA Wood. The biogenesis and characterization of mammalian microRNAs of mirtron origin. *Nucleic acids research*, 40(1):438–448, 2012. doi:10.1093/nar/gkr722.

[361] Xi Chen, Yi Ba, Lijia Ma, Xing Cai, Yuan Yin, Kehui Wang, Jigang Guo, Yujing Zhang, Jiangning Chen, Xing Guo, et al. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell research*, 18(10):997–1006, 2008. doi:10.1038/cr.2008.282.

[362] John P Cogswell, James Ward, Ian A Taylor, Michelle Waters, Yunling Shi, Brian Cannon, Kevin Kelnar, Jon Kemppainen, David Brown, Caifu Chen, et al. Identification of miRNA changes in Alzheimer's disease brain and CSF yields putative biomarkers and insights into disease pathways. *Journal of Alzheimer's disease*, 14(1):27–41, 2008. doi:10.3233/JAD-2008-14103.

[363] Alessia Gallo, Mayank Tandon, Ilias Alevizos, and Gabor G Illei. The majority of microRNAs detectable in serum and saliva is concentrated in exosomes. *PloS one*, 7(3):e30679, 2012. doi:10.1371/journal.pone.0030679.

[364] Bethany N Hannafon, Karla J Carpenter, William L Berry, Ralf Janknecht, William C Dooley, and Wei-Qun Ding. Exosome-mediated microRNA signaling from breast cancer cells is altered by the anti-angiogenesis agent docosahexaenoic acid (DHA). *Molecular cancer*, 14(1):1–13, 2015. doi:10.1186/s12943-015-0400-7.

[365] Jie Zhu, Zhibao Zheng, Jia Wang, Jinhua Sun, Pan Wang, Xianying Cheng, Lun Fu, Liming Zhang, Zuojun Wang, and Zhaoyun Li. Different miRNA expression profiles between human breast cancer tumors and serum. *Frontiers in genetics*, 5:149, 2014. doi:10.3389/fgene.2014.00149.

[366] Yin Hu, Shan-Shan Rao, Zhen-Xing Wang, Jia Cao, Yi-Juan Tan, Juan Luo, Hong-Ming Li, Wei-She Zhang, Chun-Yuan Chen, and Hui Xie. Exosomes from human umbilical cord blood accelerate cutaneous wound healing through mir-21-3p-mediated promotion of angiogenesis and fibroblast function. *Theranostics*, 8(1):169, 2018. doi:10.7150/thno.21234.

[367] Yong Peng, Yuntao Dai, Charles Hitchcock, Xiaojuan Yang, Edmund S Kassis, Lunxu Liu, Zhenghua Luo, Hui-Lung Sun, Ri Cui, Huijun Wei, et al. Insulin growth factor signaling is regulated by microRNA-486, an underexpressed microRNA in lung cancer. *Proceedings of the national academy of sciences*, 110(37):15043–15048, 2013. doi:10.1073/pnas.1307107110.

[368] Anchal Vishnoi and Sweta Rani. MiRNA biogenesis and regulation of diseases: an overview. In *MicroRNA Profiling*, pages 1–10. Springer, 2017. doi:10.1007/978-1-4939-6524-3_1.

[369] ES Maxwell and MJ Fournier. The small nucleolar RNAs. *Annual review of biochemistry*, 64(1): 897–934, 1995. doi:10.1146/annurev.bi.64.070195.004341.

[370] Junnan Liang, Jingyuan Wen, Zhao Huang, Xiao-ping Chen, Bi-xiang Zhang, and Liang Chu. Small nucleolar RNAs: insight into their function in Cancer. *Frontiers in oncology*, 9, 2019. doi:10.3389/fonc.2019.00587.

[371] Hadi Jorjani, Stephanie Kehr, Dominik J Jedlinski, Rafal Gumienny, Jana Hertel, Peter F Stadler, Mihaela Zavolan, and Andreas R Gruber. An updated human snoRNAome. *Nucleic acids research*, 44(11):5068–5082, 2016. doi:10.1093/nar/gkw386.

[372] Tamás Kiss. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, 109(2):145–148, 2002. doi:10.1016/S0092-8674(02)00718-3.

[373] Steve L Reichow, Tomoko Hamma, Adrian R Ferré-D'Amaré, and Gabriele Varani. The structure and function of small nucleolar ribonucleoproteins. *Nucleic acids research*, 35(5):1452–1464, 2007. doi:10.1093/nar/gkl1172.

[374] Tamás Kiss. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *The EMBO journal*, 20(14):3617–3622, 2001. doi:10.1093/emboj/20.14.3617.

[375] Zsuzsanna Kiss-László, Yves Henry, Jean-Pierre Bachellerie, Michèle Caizergues-Ferrer, and Tamás Kiss. Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, 85(7):1077–1088, 1996. doi:10.1016/S0092-8674(00)81308-2.

[376] Jérôme Cavaillé, Monique Nicoloso, and Jean-Pierre Bachellerie. Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides. *Nature*, 383(6602):732–735, 1996. doi:10.1038/383732a0.

[377] Kenneth Scott McKeegan, Charles Maurice Debieux, Séverine Boulon, Edouard Bertrand, and Nicholas James Watkins. A dynamic scaffold of pre-snoRNP factors facilitates human box C/D snoRNP assembly. *Molecular and cellular biology*, 27(19):6782–6793, 2007. doi:10.1128/mcb.01097-07.

[378] Zsuzsanna Kiss-László, Yves Henry, and Tamás Kiss. Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *The EMBO journal*, 17(3):797–807, 1998. doi:10.1093/emboj/17.3.797.

[379] Andrey G Balakin, Laurie Smith, and Maurille J Fournier. The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. *Cell*, 86(5):823–834, 1996. doi:10.1016/S0092-8674(00)80156-7.

[380] Susan Kass, Kazimierz Tyc, Joan A Steitz, and Barbara Sollner-Webb. The U3 small nucleolar ribonucleoprotein functions in the first step of preribosomal RNA processing. *Cell*, 60(6):897–908, 1990. doi:10.1016/0092-8674(90)90338-F.

[381] Nicholas J Watkins and Markus T Bohnsack. The box C/D and H/ACA snoRNPs: key players in the modification, processing and the dynamic folding of ribosomal RNA. *Wiley Interdisciplinary Reviews: RNA*, 3(3):397–414, 2012. doi:10.1002/wrna.117.

[382] Kazimierz T Tycowski, Zhi-Hao You, Paul J Graham, and Joan A Steitz. Modification of U6 spliceosomal RNA is guided by other small RNAs. *Molecular cell*, 2(5):629–638, 1998. doi:10.1016/S1097-2765(00)80161-6.

[383] Jérôme Cavaillé, Karin Buiting, Martin Kiefmann, Marc Lalande, Camilynn I Brannan, Bernhard Horsthemke, Jean-Pierre Bachellerie, Jürgen Brosius, and Alexander Hüttenhofer. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proceedings of the National Academy of Sciences*, 97(26):14311–14316, 2000. doi:10.1073/pnas.250426397.

[384] Patrice Vitali, Eugenia Basyuk, Elodie Le Meur, Edouard Bertrand, Françoise Muscatelli, Jérôme Cavaillé, and Alexander Huttenhofer. ADAR2-mediated editing of RNA substrates in the nucleolus is inhibited by C/D small nucleolar RNAs. *The Journal of cell biology*, 169(5): 745–753, 2005. doi:10.1083/jcb.200411129.

[385] Shivendra Kishore, Amit Khanna, Zhaiyi Zhang, Jingyi Hui, Piotr J Balwierz, Mihaela Stefan, Carol Beach, Robert D Nicholls, Mihaela Zavolan, and Stefan Stamm. The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Human molecular genetics*, 19(7):1153–1164, 2010. doi:10.1093/hmg/ddp585.

[386] Motoharu Ono, Kayo Yamada, Fabio Avolio, Michelle S Scott, Silvana van Koningsbruggen, Geoffrey J Barton, and Angus I Lamond. Analysis of human small nucleolar RNAs (snoRNA) and the development of snoRNA modulator of gene expression vectors. *Molecular biology of the cell*, 21(9):1569–1584, 2010. doi:10.1091/mbc.E10-01-0078.

[387] Liang Chu, Mack Y Su, Leonard B Maggi, Lan Lu, Chelsea Mullins, Seth Crosby, Gaofeng Huang, Wee Joo Chng, Ravi Vij, Michael H Tomasson, et al. Multiple myeloma–associated chromosomal translocation activates orphan snoRNA ACA11 to suppress oxidative stress. *The Journal of clinical investigation*, 122(8):2793–2806, 2012. doi:10.1172/JCI63051.

[388] Jean-Louis Langhendries, Emilien Nicolas, Gilles Doumont, Serge Goldman, and Denis LJ Lafontaine. The human box C/D snoRNAs U3 and U8 are required for pre-rRNA processing and tumorigenesis. *Oncotarget*, 7(37):59519, 2016. doi:10.18632/oncotarget.11148.

[389] H Su, T Xu, S Ganapathy, M Shadfan, M Long, T HM Huang, I Thompson, and ZM Yuan. Elevated snoRNA biogenesis is essential in breast cancer. *Oncogene*, 33(11):1348–1358, 2014. doi:10.1038/onc.2013.89.

[390] Lu Gao, Jie Ma, Kaiissar Mannoor, Maria A Guarnera, Amol Shetty, Min Zhan, Lingxiao Xing, Sanford A Stass, and Feng Jiang. Genome-wide small nucleolar RNA expression analysis of lung cancer by next-generation deep sequencing. *International journal of cancer*, 136(6):E623–E629, 2015. doi:10.1002/ijc.29169.

[391] Preethi Krishnan, Sunita Ghosh, Bo Wang, Mieke Heyns, Kathryn Graham, John R Mackey, Olga Kovalchuk, and Sambasivarao Damaraju. Profiling of small nucleolar RNAs by next generation sequencing: potential new players for breast cancer prognosis. *PloS one*, 11(9):e0162622, 2016. doi:10.1371/journal.pone.0162622.

[392] HE Gee, FM Buffa, C Camps, A Ramachandran, R Leek, M Taylor, M Patil, H Sheldon, G Betts, J Homer, et al. The small-nucleolar RNAs commonly used for microRNA normalisation correlate with tumour pathology and prognosis. *British journal of cancer*, 104(7):1168–1177, 2011. doi:10.1038/sj.bjc.6606076.

[393] Laure Berquet, Wilfried Valleron, Srdana Grgurevic, Cathy Quelen, Ouafa Zaki, Anne Quillet-Mary, Frederic Davi, Pierre Brousset, Marina Bousquet, and Loïc Ysebaert. Small nucleolar RNA expression profiles refine the prognostic impact of IGHV mutational status on treatment-free survival in chronic lymphocytic leukaemia. *British Journal of Haematology*, 172(5):819–823, 2015. doi:10.1111/bjh.13544.

[394] Jonathan Krell, Adam E Frampton, Reza Mirnezami, Victoria Harding, Alex De Giorgio, Laura Roca Alonso, Patrizia Cohen, Silvia Ottaviani, Teresa Colombo, Jimmy Jacob, et al. Growth arrest-specific transcript 5 associated snoRNA levels are related to p53 expression and DNA damage in colorectal cancer. *PloS one*, 9(6):e98561, 2014. doi:10.1371/journal.pone.0098561.

[395] Tomaž Bratkovič, Janja Božič, and Boris Rogelj. Functional diversity of small nucleolar RNAs. *Nucleic acids research*, 48(4):1627–1651, 2020. doi:10.1093/nar/gkz1140.

[396] Nicolas Lemus-Diaz, Rafael Rinaldi Ferreira, Katherine E Bohnsack, Jens Gruber, and Markus T Bohnsack. The human box C/D snoRNA U3 is a miRNA source and mir-U3 regulates expression of sortin nexin 27. *Nucleic acids research*, 48(14):8074–8089, 2020. doi:10.1093/nar/gkaa549.

[397] Ryan J Taft, Evgeny A Glazov, Timo Lassmann, Yoshihide Hayashizaki, Piero Carninci, and John S Mattick. Small RNAs derived from snoRNAs. *RNA*, 15(7):1233–1240, 2009. doi:10.1261/rna.1528909.

[398] Markus Brameier, Astrid Herwig, Richard Reinhardt, Lutz Walter, and Jens Gruber. Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs. *Nucleic acids research*, 39(2):675–686, 2011. doi:10.1093/nar/gkq776.

[399] Feng Yu, Cameron P Bracken, Katherine A Pillman, David M Lawrence, Gregory J Goodall, David F Callen, and Paul M Neilsen. p53 represses the oncogenic Sno-MiR-28 derived from a SnoRNA. *PloS one*, 10(6):e0129190, 2015. doi:10.1371/journal.pone.0129190.

[400] Jian-You Liao, Li-Ming Ma, Yan-Hua Guo, Yu-Chan Zhang, Hui Zhou, Peng Shao, Yue-Qin Chen, and Liang-Hu Qu. Deep sequencing of human nuclear and cytoplasmic small RNAs reveals an unexpectedly complex subcellular distribution of miRNAs and tRNA 3' trailers. *PloS one*, 5 (5):e10563, 2010. doi:10.1371/journal.pone.0010563.

[401] Motoharu Ono, Michelle S Scott, Kayo Yamada, Fabio Avolio, Geoffrey J Barton, and Angus I Lamond. Identification of human miRNA precursors that resemble box C/D snoRNAs. *Nucleic acids research*, 39(9):3879–3891, 2011. doi:10.1093/nar/gkq1355.

[402] John WS Brown, David F Marshall, and Manuel Echeverria. Intronic noncoding RNAs and splicing. *Trends in plant science*, 13(7):335–342, 2008. doi:10.1016/j.tplants.2008.04.010.

[403] Gyorgy Hutvagner and Martin J Simard. Argonaute proteins: key players in RNA silencing. *Nature reviews Molecular cell biology*, 9(1):22–32, 2008. doi:10.1038/nrm2321.

[404] Jiening Xiao, Huixian Lin, Xiaobin Luo, Xiaoyan Luo, and Zhiguo Wang. miR-605 joins p53 network to form a p53: miR-605: Mdm2 positive feedback loop in response to stress. *The EMBO journal*, 30(3):524–532, 2011. doi:10.1038/emboj.2010.347.

[405] Michelle S Scott, Fabio Avolio, Motoharu Ono, Angus I Lamond, and Geoffrey J Barton. Human miRNA precursors with box H/ACA snoRNA features. *PLoS Comput Biol*, 5(9):e1000507, 2009. doi:10.1371/journal.pcbi.1000507.

[406] Athanasius F Bompfünewerer, Christoph Flamm, Claudia Fried, Guido Fritzsch, Ivo L Hofacker, Jörg Lehmann, Kristin Missal, Axel Mosig, Bettina Müller, Sonja J Prohaska, et al. Evolutionary patterns of non-coding RNAs. *Theory in Biosciences*, 123(4):301–369, 2005. doi:10.1016/j.thbio.2005.01.002.

[407] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. miRBase: from microRNA sequences to function. *Nucleic acids research*, 47(D1):D155–D162, 2019. doi:10.1093/nar/gky1141.

[408] Philia Bouchard-Bourelle, Clément Desjardins-Henri, Darren Mathurin-St-Pierre, Gabrielle Deschamps-Francoeur, Étienne Fafard-Couture, Jean-Michel Garant, Sherif Abou Elela, and Michelle S Scott. snoDB: an interactive database of human snoRNA sequences, abundance and interactions. *Nucleic acids research*, 48(D1):D220–D225, 2020. doi:10.1093/nar/gkz884.

[409] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010. doi:10.1093/bioinformatics/btq033.

[410] Jeffrey M Perkel. Why Jupyter is data scientists' computational notebook of choice. *Nature*, 563 (7732):145–147, 2018. doi:10.1038/d41586-018-07196-1.

[411] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian E Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica B Hamrick, Jason Grout, Sylvain Corlay, et al. Jupyter Notebooks-a publishing format for reproducible computational workflows. In *ELPUB*, pages 87–90, 2016. doi:10.3233/978-1-61499-649-1-87.

[412] The pandas development team. pandas-dev/pandas: Pandas, February 2020. URL `https://doi.org/10.5281/zenodo.3509134`.

[413] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi:10.25080/Majora-92bf1922-00a.

[414] Charles R. Harris, K. Jarrod Millman, Stefan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernandez del Rio, Mark Wiebe, Pearu Peterson, Pierre Gerard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi:10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

[415] François Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[416] Guillaume Marçais and Carl Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011. doi:10.1093/bioinformatics/btr011.

[417] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome research*, 12(6): 996–1006, 2002. doi:10.1101/gr.229102.

[418] Mario Fasold, David Langenberger, Hans Binder, Peter F Stadler, and Steve Hoffmann. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic acids research*, 39(suppl_2):W112–W117, 2011. doi:10.1093/nar/gkr357.

[419] Müşerref Duygu Saçar and Jens Allmer. Machine learning methods for microRNA gene prediction. In *miRNomics: MicroRNA Biology and Computational Analysis*, pages 177–187. Springer, 2014. doi:10.1007/978-1-62703-748-8_10.

[420] David Langenberger, Sebastian Bartschat, Jana Hertel, Steve Hoffmann, Hakim Tafer, and Peter F Stadler. MicroRNA or not microRNA? In *Brazilian Symposium on Bioinformatics*, pages 1–9. Springer, 2011. doi:10.1007/978-3-642-22825-4_1.

[421] Lily Agranat-Tamir, Noam Shomron, Joseph Sperling, and Ruth Sperling. Interplay between pre-mRNA splicing and microRNA biogenesis within the supraspliceosome. *Nucleic acids research*, 42(7):4640–4651, 2014. doi:10.1093/nar/gkt1413.

[422] Giulia Pianigiani, Danilo Licastro, Paola Fortugno, Daniele Castiglia, Ivana Petrovic, and Franco Pagani. Microprocessor-dependent processing of splice site overlapping microRNA exons does not result in changes in alternative splicing. *RNA*, 24(9):1158–1171, 2018. doi:10.1261/rna.063438.117.

[423] Tetsuro Hirose, Mei-Di Shu, and Joan A Steitz. Splicing-dependent and-independent modes of assembly for intron-encoded box C/D snoRNPs in mammalian cells. *Molecular cell*, 12(1): 113–123, 2003. doi:10.1016/S1097-2765(03)00267-3.

[424] Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994. doi:10.1007/BF00818163.

[425] Rolf Backofen, Jan Gorodkin, Ivo L Hofacker, and Peter F Stadler. Comparative RNA genomics. In *Comparative Genomics*, pages 363–400. Springer, 2018. doi:10.1007/978-1-4939-7463-4_14.

[426] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4:e05005, 2015. doi:10.7554/eLife.05005.

[427] Dimitra Karagkouni, Maria D Paraskevopoulou, Spyros Tastsoglou, Giorgos Skoufos, Anna Karavangeli, Vasilis Pierros, Elissavet Zacharopoulou, and Artemis G Hatzigeorgiou. DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts. *Nucleic acids research*, 48(D1):D101–D110, 2020. doi:10.1093/nar/gkz1036.

[428] Eugene Berezikov. Evolution of microRNA diversity and regulation in animals. *Nature Reviews Genetics*, 12(12):846–860, 2011. doi:10.1038/nrg3079.

[429] Jana Hertel and Peter F Stadler. The expansion of animal microRNA families revisited. *Life*, 5 (1):905–920, 2015. doi:10.3390/life5010905.

[430] Yehu Moran, Maayan Agron, Daniela Praher, and Ulrich Technau. The evolutionary origin of plant and animal microRNAs. *Nature ecology & evolution*, 1(3):1–8, 2017. doi:10.1038/s41559-016-0027.

[431] Stephanie Kehr, Sebastian Bartschat, Hakim Tafer, Peter F Stadler, and Jana Hertel. Matching of Soulmates: coevolution of snoRNAs and their targets. *Molecular biology and evolution*, 31(2): 455–467, 2014. doi:10.1093/molbev/mst209.

[432] Anne Nitsche, Dominic Rose, Mario Fasold, Kristin Reiche, and Peter F Stadler. Comparison of splice sites reveals that long noncoding RNAs are evolutionarily well conserved. *RNA*, 21(5): 801–812, 2015. doi:10.1261/rna.046342.114.

[433] Andrzej T Wierzbicki. The role of long non-coding RNA in transcriptional gene silencing. *Current opinion in plant biology*, 15(5):517–522, 2012. doi:10.1016/j.pbi.2012.08.008.

[434] Urminder Singh, Niraj Khemka, Mohan Singh Rajkumar, Rohini Garg, and Mukesh Jain. PLncPRO for prediction of long non-coding RNAs (lncRNAs) in plants and its application for discovery of abiotic stress-responsive lncRNAs in rice and chickpea. *Nucleic acids research*, 45 (22):e183–e183, 2017. doi:10.1093/nar/gkx866.

[435] S. Zhang, F. Zheng, L. Zhang, Z. Huang, X. Huang, Z. Pan, S. Chen, C. Xu, Y. Jiang, S. Gu, C. Zhao, Q. Zhang, and G. Shi. LncRNA HOTAIR-mediated MTHFR methylation inhibits 5-fluorouracil sensitivity in esophageal cancer cells. *J Exp Clin Cancer Res*, 39(1):131, Jul 2020. doi:10.1186/s13046-020-01610-1.

[436] C. Poulet, M. S. Njock, C. Moermans, E. Louis, R. Louis, M. Malaise, and J. Guiot. Exosomal Long Non-Coding RNAs in Lung Diseases. *Int J Mol Sci*, 21(10), May 2020. doi:10.3390/ijms21103580.

[437] Y. H. Zhou, Y. H. Cui, T. Wang, and Y. Luo. Long non-coding RNA HOTAIR in cervical cancer: Molecular marker, mechanistic insight, and therapeutic target. *Adv Clin Chem*, 97: 117–140, 2020. doi:10.1016/bs.acc.2019.12.004.

[438] D. Jiang, L. Xu, J. Ni, J. Zhang, M. Cai, and L. Shen. Functional polymorphisms in LncRNA HOTAIR contribute to susceptibility of pancreatic cancer. *Cancer Cell Int*, 19:47, 2019.

[439] L. Qian, Q. Fei, H. Zhang, M. Qiu, B. Zhang, Q. Wang, Y. Yu, C. Guo, Y. Ren, M. Mei, L. Zhang, Y. Zhu, and B. Yang. lncRNA HOTAIR Promotes DNA Repair and Radioresistance of Breast Cancer via EZH2. *DNA Cell Biol*, Nov 2020. doi:10.1186/s12935-019-0761-x.

[440] D. Tao, Z. Zhang, X. Liu, Z. Zhang, Y. Fu, P. Zhang, H. Yuan, L. Liu, J. Cheng, and H. Jiang. LncRNA HOTAIR promotes the invasion and metastasis of oral squamous cell carcinoma through metastasis-associated gene 2. *Mol Carcinog*, 59(4):353–364, 04 2020. doi:10.1002/mc.23159.

[441] X. Gong and Z. Zhu. Long Noncoding RNA HOTAIR Contributes to Progression in Hepatocellular Carcinoma by Sponging miR-217-5p. *Cancer Biother Radiopharm*, 35(5):387–396, Jun 2020. doi:10.1089/cbr.2019.3070.

[442] J. Shengnan, X. Dafei, J. Hua, F. Sunfu, W. Xiaowei, and X. Liang. Long non-coding RNA HOTAIR as a competitive endogenous RNA to sponge miR-206 to promote colorectal cancer progression by activating CCL2. *J Cancer*, 11(15):4431–4441, 2020. doi:10.7150/jca.42308.

[443] M. P. Sperandeo, P. Ungaro, M. Vernucci, P. V. Pedone, F. Cerrato, L. Perone, S. Casola, M. V. Cubellis, C. B. Bruni, G. Andria, G. Sebastio, and A. Riccio. Relaxation of insulin-like growth factor 2 imprinting and discordant methylation at KvDMR1 in two first cousins affected by Beckwith-Wiedemann and Klippel-Trenaunay-Weber syndromes. *Am J Hum Genet*, 66(3): 841–847, Mar 2000. doi:10.1086/302811.

[444] D. Rovina, M. La Vecchia, A. Cortesi, L. Fontana, M. Pesant, S. Maitz, S. Tabano, B. Bodega, M. Miozzo, and S. M. Sirchia. Profound alterations of the chromatin architecture at chromosome 11p15.5 in cells from Beckwith-Wiedemann and Silver-Russell syndromes patients. *Sci Rep*, 10 (1):8275, 05 2020. doi:10.1038/s41598-020-65082-1.

[445] H. Cheng, H. Zhao, X. Xiao, Q. Huang, W. Zeng, B. Tian, T. Ma, D. Lu, Y. Jin, and Y. Li. Long Non-coding RNA MALAT1 Upregulates ZEB2 Expression to Promote Malignant Progression of Glioma by Attenuating miR-124. *Mol Neurobiol*, Oct 2020. doi:10.1007/s12035-020-02165-0.

[446] S. Ghafouri-Fard, M. Esmaeili, M. Taheri, and M. Samsami. Highly upregulated in liver cancer (HULC): An update on its role in carcinogenesis. *J Cell Physiol*, 235(12):9071–9079, Dec 2020. doi:10.1002/jcp.29765.

[447] X. Liu, W. Shang, and F. Zheng. Long non-coding RNA NEAT1 promotes migration and invasion of oral squamous cell carcinoma cells by sponging microRNA-365. *Exp Ther Med*, 16(3): 2243–2250, Sep 2018. doi:10.3892/etm.2018.6493.

[448] X. Zhou, X. Wang, Y. Zhou, L. Cheng, Y. Zhang, and Y. Zhang. Long Noncoding RNA NEAT1 Promotes Cell Proliferation And Invasion And Suppresses Apoptosis In Hepatocellular Carcinoma By Regulating miRNA-22-3p/akt2 In Vitro And In Vivo. *Onco Targets Ther*, 12:8991–9004, 2019. doi:10.2147/OTT.S224521.

[449] A. K. Murugan, A. K. Munirajan, and A. S. Alzahrani. Long noncoding RNAs: emerging players in thyroid cancer pathogenesis. *Endocr Relat Cancer*, 25(2):R59–R82, 02 2018. doi:10.1530/ERC-17-0188.

[450] S. Zhao, N. F. Fan, X. H. Chen, C. H. Zhuo, C. W. Xu, and R. B. Lin. Long noncoding RNA PVT1-214 enhances gastric cancer progression by upregulating TrkC expression in competitively sponging way. *Eur Rev Med Pharmacol Sci*, 24(16):8245, Aug 2020. doi:10.26355/eurrev_201905_17920.

[451] Á. Martínez-Barriocanal, D. Arango, and H. Dopeso. PVT1 Long Non-coding RNA in Gastrointestinal Cancer. *Front Oncol*, 10:38, 2020.

[452] R. Xu, Y. Mao, K. Chen, W. He, W. Shi, and Y. Han. The long noncoding RNA ANRIL acts as an oncogene and contributes to paclitaxel resistance of lung adenocarcinoma A549 cells. *Oncotarget*, 8(24):39177–39184, Jun 2017. doi:10.3389/fonc.2020.00038.

[453] M. D. Huang, W. M. Chen, F. Z. Qi, R. Xia, M. Sun, T. P. Xu, L. Yin, E. B. Zhang, W. De, and Y. Q. Shu. Long non-coding RNA ANRIL is upregulated in hepatocellular carcinoma and regulates cell apoptosis by epigenetic silencing of KLF2. *J Hematol Oncol*, 8:50, May 2015. doi:10.1186/s13045-015-0146-0.

[454] M. J. Hoffmann, J. Dehn, J. Droop, G. Niegisch, C. Niedworok, T. Szarvas, and W. A. Schulz. Truncated Isoforms of lncRNA ANRIL Are Overexpressed in Bladder Cancer, But Do Not Contribute to Repression of INK4 Tumor Suppressors. *Noncoding RNA*, 1(3):266–284, Dec 2015. doi:10.3390/ncrna1030266.

[455] S. Ghafouri-Fard, S. Dashti, and M. Taheri. PCAT1: An oncogenic lncRNA in diverse cancers and a putative therapeutic target. *Exp Mol Pathol*, 114:104429, 06 2020. doi:10.1016/j.yexmp.2020.104429.

[456] M. Shademan, A. Naseri Salanghuch, K. Zare, M. Zahedi, M. A. Foroughi, K. Akhavan Rezayat, H. Mosannen Mozaffari, K. Ghaffarzadegan, L. Goshayeshi, and H. Dehghani. Expression profile analysis of two antisense lncRNAs to improve prognosis prediction of colorectal adenocarcinoma. *Cancer Cell Int*, 19:278, 2019. doi:10.1186/s12935-019-1000-1.

[457] G. D. Hu, C. X. Wang, H. Y. Wang, Y. Q. Wang, S. Hu, Z. W. Cao, B. Min, L. Li, X. F. Tian, and H. B. Hu. Long noncoding RNA CCAT2 functions as a competitive endogenous RNA to regulate FOXC1 expression by sponging miR-23b-5p in lung adenocarcinoma. *J Cell Biochem*, Dec 2018. doi:10.1002/jcb.28077.

[458] H. Ling, R. Spizzo, Y. Atlasi, M. Nicoloso, M. Shimizu, R. S. Redis, N. Nishida, R. Gaf?, J. Song, Z. Guo, C. Ivan, E. Barbarotto, I. De Vries, X. Zhang, M. Ferracin, M. Churchman, J. F. van Galen, B. H. Beverloo, M. Shariati, F. Haderk, M. R. Estecio, G. Garcia-Manero, G. A. Patijn, D. C. Gotley, V. Bhardwaj, I. Shureiqi, S. Sen, A. S. Multani, J. Welsh, K. Yamamoto, I. Taniguchi, M. A. Song, S. Gallinger, G. Casey, S. N. Thibodeau, L. Le Marchand, M. Tiirikainen, S. A. Mani, W. Zhang, R. V. Davuluri, K. Mimori, M. Mori, A. M. Sieuwerts, J. W. Martens, I. Tomlinson, M. Negrini, I. Berindan-Neagoe, J. A. Foekens, S. R. Hamilton, G. Lanza, S. Kopetz, R. Fodde, and G. A. Calin. CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. *Genome Res*, 23(9):1446–1461, Sep 2013. doi:10.1101/gr.152942.112.

[459] J. Dong, M. Su, W. Chang, K. Zhang, S. Wu, and T. Xu. Long non-coding RNAs on the stage of cervical cancer (Review). *Oncol Rep*, 38(4):1923–1931, Oct 2017. doi:10.3892/or.2017.5905.

[460] N. F. Hosseini, H. Manoochehri, S. G. Khoei, and M. Sheykhhasan. The Functional Role of Long Non-coding RNA UCA1 in Human Multiple Cancers: a Review Study. *Curr Mol Med*, Jun 2020. doi:10.2174/1566524020666200619124543.

[461] Z. Zeng, H. Bo, Z. Gong, Y. Lian, X. Li, X. Li, W. Zhang, H. Deng, M. Zhou, S. Peng, G. Li, and W. Xiong. AFAP1-AS1, a long noncoding RNA upregulated in lung cancer and promotes invasion and metastasis. *Tumour Biol*, 37(1):729–737, Jan 2016. doi:10.1007/s13277-015-3860-x.

[462] D. Ji, X. Zhong, X. Jiang, K. Leng, Y. Xu, Z. Li, L. Huang, J. Li, and Y. Cui. The role of long non-coding RNA AFAP1-AS1 in human malignant tumors. *Pathol Res Pract*, 214(10): 1524–1531, Oct 2018. doi:10.1016/j.prp.2018.08.014.

[463] M. Barton, J. Santucci-Pereira, O. G. Vaccaro, T. Nguyen, Y. Su, and J. Russo. BC200 overexpression contributes to luminal and triple negative breast cancer pathogenesis. *BMC Cancer*, 19(1):994, Oct 2019. doi:10.1186/s12885-019-6179-y.

[464] H. Shin, J. Lee, Y. Kim, S. Jang, T. Ohn, and Y. Lee. Identifying the cellular location of brain cytoplasmic 200 RNA using an RNA-recognizing antibody. *BMB Rep*, 50(6):318–322, Jun 2017. doi:10.5483/BMBRep.2017.50.6.217.

[465] L. Poliseno, L. Salmena, J. Zhang, B. Carver, W. J. Haveman, and P. P. Pandolfi. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465 (7301):1033–1038, Jun 2010. doi:10.1038/nature09144.

[466] J. Mazar, W. Zhao, A. M. Khalil, B. Lee, J. Shelley, S. S. Govindarajan, F. Yamamoto, M. Ratnam, M. N. Aftab, S. Collins, B. N. Finck, X. Han, J. S. Mattick, M. E. Dinger, and R. J. Perera. The functional characterization of long noncoding RNA SPRY4-IT1 in human melanoma cells. *Oncotarget*, 5(19):8959–8969, Oct 2014. doi:10.18632/oncotarget.1863.

[467] W. Feng, C. Wang, C. Liang, H. Yang, D. Chen, X. Yu, W. Zhao, D. Geng, S. Li, Z. Chen, and M. Sun. The Dysregulated Expression of KCNQ1OT1 and Its Interaction with Downstream Factors miR-145/CCNE2 in Breast Cancer Cells. *Cell Physiol Biochem*, 49(2):432–446, 2018. doi:10.1159/000492978.

[468] Y. Wang, L. Zhang, J. Yang, and R. Sun. LncRNA KCNQ1OT1 promotes cell proliferation, migration and invasion via regulating miR-129-5p/JAG1 axis in non-small cell lung cancer. *Cancer Cell Int*, 20:144, 2020. doi:10.1186/s12935-020-01225-8.

[469] A. Poursheikhani, M. R. Abbaszadegan, N. Nokhandani, and M. A. Kerachian. Integration analysis of long non-coding RNA (lncRNA) role in tumorigenesis of colon adenocarcinoma. *BMC Med Genomics*, 13(1):108, 07 2020. doi:10.1186/s12920-020-00757-2.

[470] Y. Liao, B. Zhang, T. Zhang, Y. Zhang, and F. Wang. LncRNA GATA6-AS Promotes Cancer Cell Proliferation and Inhibits Apoptosis in Glioma by Downregulating lncRNA TUG1. *Cancer Biother Radiopharm*, 34(10):660–665, Dec 2019. doi:10.1089/cbr.2019.2830.

[471] A. Al-Rugeebah, M. Alanazi, and N. R. Parine. MEG3: an Oncogenic Long Non-coding RNA in Different Cancers. *Pathol Oncol Res*, 25(3):859–874, Jul 2019. doi:10.1007/s12253-019-00614-3.

[472] A. He, S. He, X. Li, and L. Zhou. ZFAS1: A novel vital oncogenic lncRNA in multiple human cancers. *Cell Prolif*, 52(1):e12513, Jan 2019. doi:10.1111/cpr.12513.

[473] Y. Zhang, L. J. Fan, Y. Zhang, J. Jiang, and X. W. Qi. Long Non-coding Wilms Tumor 1 Antisense RNA in the Development and Progression of Malignant Tumors. *Front Oncol*, 10:35, 2020. doi:10.3389/fonc.2020.00035.

[474] A. Touati, J. Errea-Dorronsoro, S. Nouri, Y. Halleb, A. Pereda, N. Mahdhaoui, A. Ghith, A. Saad, G. Perez de Nanclares, and D. H'mida Ben Brahim. Transient neonatal diabetes mellitus and hypomethylation at additional imprinted loci: novel ZFP57 mutation and review on the literature. *Acta Diabetol*, 56(3):301–307, Mar 2019. doi:10.1007/s00592-018-1239-3.

[475] Sébastien Bonnet, Olivier Boucherat, Roxane Paulin, Danchen Wu, Charles CT Hindmarch, Stephen L Archer, Rui Song, Joseph B Moore IV, Steeve Provencher, Lubo Zhang, et al. Clinical value of non-coding RNAs in cardiovascular, pulmonary, and muscle diseases. *American Journal of Physiology-Cell Physiology*, 318(1):C1–C28, 2020. doi:10.1152/ajpcell.00078.2019.

[476] A. Suwal, J. L. Hao, X. F. Liu, D. D. Zhou, O. P. Pant, Y. Gao, P. Hui, X. X. Dai, and C. W. Lu. NONRATT021972 long-noncoding RNA: A promising lncRNA in diabetes-related diseases. *Int J Med Sci*, 16(6):902–908, 2019. doi:10.7150/ijms.34200.

# Curriculum Scientiae

## Personal Information

| | |
|---|---|
| Name | Rituparno Sen |
| Date of birth | $17^{th}$ of May, 1988 |
| Place of birth | Coochbehar, India |

## Education

| | |
|---|---|
| Oct, 2015 : Mar, 2021 | PhD Student<br>Bioinformatics Group,<br>University of Leipzig, Germany |
| Aug, 2009 : May, 2011 | M.Sc. Student<br>Department of Computer Science,<br>Pondicherry University, India<br>Title of Master's Thesis: Intrusion Detection System using Genetic Algorithms |
| June, 2006 : April, 2009 | B.Sc. Student<br>Department of Computer Science,<br>St. Xavier's College, Kolkata, India |

## Awards

2015 Research Grant for Doctoral Programme, (2015-2020)
Awarded by the Deutscher Akademischer Austausch Dienst, Germany

## Research Experience

| | |
|---|---|
| 2012 : 2015 | Graduate Research Assistant<br>Indian Association for the Cultivation of Science, Kolkata<br>and Gyanxet<br>Small regulatory RNAs<br>Studying interaction of human long non-coding RNAs<br>with mRNAs and miRNAs and their disease associations |

## Technical Knowledge

| | |
|---|---|
| Programming Languages | Python, Java, R, C++, Shell |
| Concepts | Bioinformatic Tools, RNA Sequencing, Machine Learning |

## Languages

| | |
|---|---|
| English | Full professional proficiency |
| German | Professional working proficiency |
| Hindi | Professional working proficieny |
| Bengali | Native speaker |

# List of publications

**Rituparno Sen**, Jörg Fallmann, Maria Emília M. T. Walter, and Peter F. Stadler. *Are spliced ncRNA host genes distinct classes of lncRNAs?* Theory in Biosciences, 2020. https://doi.org/10.1007/s12064-020-00330-6

**Rituparno Sen**, Gero Doose, and Peter F. Stadler. *Rare splice variants in long non-coding RNAs.* Non-coding RNA, 3(3):23, 2017. https://doi.org/10.3390/ncrna3030023

Suman Ghosal, Shekhar Saha, Shaoli Das, **Rituparno Sen**, Swagata Goswami, Siddhartha S. Jana, and Jayprokas Chakrabarti. *miRepress: modelling gene expression regulation by microRNA with non-conventional binding sites.* Scientific reports 6, 22334, 2016. https://doi.org/10.1038/srep22334

Suman Ghosal, Shaoli Das, **Rituparno Sen**, and Jayprokas Chakrabarti. *HumanViCe: host ceRNA network in virus infected cells in human.* Frontiers in Genetics, 5:249, 2014. https://doi.org//10.3389/fgene.2014.00249

Shaoli Das, Suman Ghosal, **Rituparno Sen**, and Jayprokas Chakrabarti. *lnCeDB: Database of Human Long Noncoding RNA Acting as Competing Endogenous RNA.* PLOS ONE 9(6): e98965, 2014. https://doi.org/10.1371/journal.pone.0098965

**Rituparno Sen**, Suman Ghosal, Shaoli Das, Subrata Balti, and Jayprokas Chakrabarti. *Competing Endogenous RNA: The Key to Posttranscriptional Regulation.* The Scientific World Journal, vol. 2014, Article ID 896206, 6 pages, 2014. https://doi.org/10.1155/2014/896206

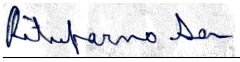
Suman Ghosal, Shaoli Das, **Rituparno Sen**, Piyali Basak, and Jayprokas Chakrabarti. *Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits.* Frontiers in Genetics, 4:283, 2013. https://doi.org//10.3389/fgene.2013.00283

# Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, 12.03.2021

(Ort, Datum)



(Unterschrift)